UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
CURSO DE ENGENHARIA DE TRANSPORTES E LOGÍSTICA

FELIPE SOARES TIBURCIO

UMA APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS PARA PREVISÃO DA DEMANDA DE PASSAGEIROS NO TRANSPORTE PÚBLICO DA CIDADE DE JOINVILLE

Joinville
2018

FELIPE SOARES TIBURCIO

UMA APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS PARA PREVISÃO DA DEMANDA DE PASSAGEIROS NO TRANSPORTE PÚBLICO DA CIDADE DE JOINVILLE

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Bacharel em Engenharia de Transportes e Logística, no curso Engenharia de Transportes e Logística da Universidade Federal de Santa Catarina, Centro Tecnológico de Joinville.

Orientador: Prof. Dr. Pablo Andretta Jaskowiak

Joinville
2018

RESUMO

Conhecer a demanda de passageiros é um aspecto fundamental do planejamento operacional de um sistema de transporte coletivo. Com o advento de Sistemas Inteligentes de Transporte e a aquisição de quantidades massivas de dados, técnicas de análise tradicionais não se mostram suficientes. Uma alternativa, nesse cenário, é o uso de técnicas provenientes do Aprendizado de Máquina e da Mineração de Dados. Inseridas neste contexto, encontram-se as Redes Neurais Artificias (RNAs), que têm sido aplicadas com sucesso à uma vasta gama de problemas. O presente trabalho investiga a aplicação de RNAs para a previsão de demanda de passageiros por dia em uma linha específica do transporte público por ônibus da cidade de Joinville - SC. Dentre os modelos avaliados, os melhores resultados obtidos indicam um erro médio percentual absoluto (EMPA) próximo à 11%.

Palavras-chave: Previsão de demanda de passageiro. Redes Neurais Articiais. Transporte Público.

ABSTRACT

Knowing passenger demand is a crucial aspect of public transportation systems operational planning. With the advent of Intelligent Transportation Systems and the acquisition of massive amounts of data, traditional analysis techniques are not sufficient. An alternative in this scenario is the use of techniques from Machine Learning and Data Mining. Inserted in this context are Artificial Neural Networks (ANNs), which have been successfully applied to a wide range of problems. The present work investigates the application of ANNs for daily passenger demand forecasting in a specific public transportation bus route in the city of Joinville - SC. Among the models evaluated, the best results obtained indicates mean absolute percentage error (MAPE) close to 11%.

Keywords: Passenger demand forecasting. Artificial Neural Networks. Public Transportation

AGRADECIMENTOS

Agradeço a Jesus Cristo pela Vida.
Agradeço a minha Esposa Leilane Belli Tiburcio por ser minha força.
Agradeço a minha Mãe Simone Tiburcio e ao meu Pai Edivaldo da Costa Tiburcio por terem me dado condições de chegar até aqui.
Agradeço a minha Irmã Fabrina Tiburcio pelo companherismo.
Agradeço a Universidade Federal de Santa Catarina, especialmente ao Corpo Docente e Técnico Administrativo pelo excelente serviço prestado a Comunidade.
Agradeço ao meu Orientador Prof. Dr. Pablo Andretta Jaskowiak pelo direcionamento.
Agradeço a Empresa Gidion Transportes e Turismo LTDA e a Empresa Passebus por fornecer os dados utilizados neste trabalho.
Agradeço a minha Sogra Maria Belli e ao meu Sogro Ivo Belli, bem como as minhas cunhadas Loriane e Liliane pelo apoio.
Agradeço aos meus familiares, em especial a minhas tias, Eduvirges e Joana Darc por terem desenpenhado papel fundamental na minha alfabetização.
Agradeço aos meus amigos, companheiros de todas as horas, sempre dispostos a me estender a mão.
Um Homem não é nada sem Deus, sem Família e sem amigos, sou grato por tudo isso.

Ser sábio é melhor do que ser forte; o conhecimento é mais importante do que a força. Afinal, antes de entrar numa batalha, é preciso planejar bem, e, quando há muitos conselheiros, é mais fácil vencer.

Bíblia Sagrada - Provérbios 24:5-6

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O transporte público é um direito de todo cidadão garantido por Lei através do Inciso XX do Artigo 21º da Constituição Federal Brasileira de 1988 (BRASIL, 1988) e da Lei Federal 12.578 de 2012 que trata da Política Nacional de Mobilidade Sustentável (BRASIL, 2012). Neste sentido, o serviço de transporte coletivo de passageiros tem como função primária ser promotor de igualdade e inclusão social, possibilitando o acesso às infraestruturas urbanas de saúde, educação, trabalho, e lazer ao cidadão.

Adicionalmente, além de ser um benefício social, o uso do transporte coletivo, que na maioria das cidades brasileiras é realizado de modo rodoviário por meio de ônibus, se torna indispensável na tentativa de minimizar os efeitos negativos ambientais e econômicos causados pelo uso excessivo de automóveis e motocicletas (EUROPEAN COMISSION, 2013; DEPARTMENT OF INFRASTRUCTURE AND REGIONAL DEVELOPMENT, 2014; BRASIL, 2013).

Evans e Lindsay (2014) argumentam que um usuário escolherá, dentre diferentes opções de transporte, aquela que lhe oferece a maior qualidade percebida (nível de serviço) ao menor custo generalizado, em outras palavras, a melhor relação de custo benefício. Deste modo, objetivando aumentar a popularidade do sistema de transporte, uma das formas que as operadores possuem de, simultaneamente, reduzir custos operacionais e aumentar o nível de serviço, é através do adequado atendimento da demanda de passageiros.

A quantidade de vagas ofertadas por um sistema de transporte por ônibus pode ser ajustada variando-se a frequência de atendimento ou optando-se por veículos de maior capacidade. Entretanto, a demanda de passageiros é independente da operação e variável ao longo do tempo. Sendo assim, por conta da natureza variável da demanda, obter uma previsão próxima do valor de fato observado para este parâmetro passa a ser um dos desafios que o operador de sistemas de transporte deve superar a fim de oferecer adequado nível de serviço e obter baixo custo de operação.

A modelagem da previsão de demanda para deslocamento de pessoas vem a longo tempo sendo dominada pelo Modelo das 4 etapas, onde a frequência de viagens produzidas e atraídas por cada zona de uma área de estudo é determinada na primeira etapa (Geração de Viagens) através de modelos matemáticos que relacionam as viagens com características demográficas, de uso do solo e indicadores socioeconômicos das zonas (ORTÚZAR; WILLUMSEN, 2011). Nas etapas subsequentes, a frequência de viagens é então utilizada como base para a produção de

tabelas de viagens que representam outros aspectos da demanda de movimentação de pessoas. Mahrsi (2017) argumenta que este modelo toma como base, essencialmente, dados de pesquisas para sua calibração, como, por exemplo, aquelas realizadas na origem de uma viagem, isto é, no domicílio, e aquelas baseadas no motivo de cada viagem.

Com os recentes avanços tecnológicos e o desenvolvimento da área de Sistemas Inteligentes de Transportes, dados de deslocamento de pessoas e cargas podem ser coletados por outras fontes, como sistemas de *Global Position Systems* (GPS), semáforos com detectores de veículos, sinais de *smartphones* e bilhetagem eletrônica do transporte público. Estes dados podem ser utilizados a fim de embasar estudos de transportes. As informações presentes nos sistemas de bilhetagem eletrônica do transporte público, em específico, podem ser utilizadas para complementar as pesquisas domiciliares realizadas no Modelo das 4 etapas e prover dados mais confiáveis para a previsão de demanda de passageiros (MAHRSI, 2017).

Se por um lado a quantidade massiva de dados disponíveis inviabiliza análises manuais, por outro, torna possível a adoção de métodos computacionais provenientes do Aprendizado de Máquina e da Mineração de Dados (TAN; STEINBACH; KUMAR, 2005) que, a partir de um conjunto de dados com demanda observada, obtidos a partir de sistemas de bilhetagem, podem vir a permitir a construção de modelos para previsão de demanda futura.

Tendo em vista este cenário, o presente trabalho visa estimar demanda de passageiros com base em algoritmos de Mineração de Dados e Aprendizado de Máquina. A justificativa para empregar esta abordagem se deve ao fato que os dados obtidos a partir do sistema de bilhetagem podem ser complementados com, por exemplo, informações climáticas, dados a respeito do dia em que a viagem foi feita, dentre outros. Espera-se que o uso dos dados, combinado a este tipo de informações leve a resultados mais precisos de estimação de demanda.

1.1 Objetivos

Abaixo são descritos o objetivo geral e os objetivos específicos deste trabalho.

1.1.1 Objetivo Geral

Avaliar a aplicação de métodos de Aprendizado de Máquina e Mineração de Dados, em específico Redes Neurais Artificiais, como modelos de produção (previsão) de viagens para auxílio no planejamento operacional do transporte público por ônibus.

1.1.2 Objetivos Específicos

A. Realizar a aquisição e pré-processamento de dados de diferentes fontes com o intuito de preparar uma base de dados suscetível à aplicação de Redes Neurais Artificiais;
B. Aplicar Redes Neurais Artificiais à base de dados pré-processada para estimar a quantidade de passageiros por dia em uma Linha específica do transporte coletivo da Cidade de Joinville e avaliar os resultados obtidos.

1.2 Organização do Trabalho

O Capítulo 2 tem por objetivo apresentar o referencial teórico das áreas de Planejamento de Transportes, Mineração de Dados e Redes Neurais. O Capítulo 3 apresenta a linha de transporte coletivo utilizada para o estudo de caso, bem como a organização do experimento como um todo, desde o pré-processamento da base de dados até as arquiteturas das RNAs utilizadas. O Capítulo 4 discute os resultados obtidos com a aplicação de RNAs como preditores da quantidade de passageiros por dia. Por último, o Capítulo 5 apresenta as conclusões sobre o experimento e sugestões de trabalhos futuros.

2 REVISÃO TEÓRICA

2.1 Aspectos do Planejamento de Transportes: Importância, Desafios e Ferramentas

Pessoas se movimentam por diversos motivos: trabalho, estudo, saúde e lazer, utilizando-se de diferentes modais de transporte. A mesma observação é válida para o deslocamento de cargas. Raramente, o ato de se locomover será um fim em si mesmo; o transporte é um meio de se alcançar outro objetivo em termos espaciais (ORTÚZAR; WILLUMSEN, 2011). Esta última afirmação resulta em duas observações. Primeiro, por ser o transporte uma atividade *secundária*, uma pessoa ao realizar um deslocamento procura minimizar os custos associados a esta atividade e, simultaneamente, maximizar o nível de serviço experimentado (EVANS; LINDSAY, 2014). Segundo, por ser o transporte dependente do espaço, a forma como as atividades de trabalho, estudo e lazer estão organizadas geograficamente determinam a direção em que os deslocamentos ocorrem (ORTÚZAR; WILLUMSEN, 2011).

O aumento da população ao longo das últimas décadas e, consequentemente, o aumento da demanda por deslocamentos, associados à falta de planejamento do espaço urbano acarretou no desequilíbrio entre a oferta de transportes e a demanda por eles. Como consequência dessa disparidade, os problemas com congestionamento, atraso nas viagens, acidentes de trânsito e poluição atingiram níveis muito além do que seriam considerados aceitáveis (ORTÚZAR; WILLUMSEN, 2011). Contudo, Ferraz e Torres (2004), Ortúzar e Willumsen (2011), Campos (2013), Vuchic (2005), compartilham da ideia que investimento de recursos, planejamento estratégico e operacional, e a implantação de sistemas e infraestruturas de transporte, possuem a capacidade de solucionar a assimetria entre a demanda e a oferta de transportes juntamente com seus efeitos negativos.

O investimento de recursos em sistemas e infraestruturas de transporte diz respeito à políticas públicas e não faz parte do escopo desse trabalho discutí-las. Sendo assim, o interesse desta seção é discutir aspectos da fase de planejamento de sistemas de transportes, especificamente, como o estudo da demanda por transporte é importante para todos os níveis de planejamento: do nível estratégico ao nível operacional.

Evans e Lindsay (2014) definem planejamento como sendo um conjunto de ações com a finalidade de alcançar um determinado objetivo. Dependendo do horizonte

temporal deste objetivo, o planejamento pode ser classificado em três níveis: estratégico, tático e operacional. O planejamento estratégico diz respeito às ações que serão tomadas para alcançar objetivos de longo prazo. Em termos de trasporte, um exemplo de objetivo de longo prazo seria a modificação substancial da participação do modal ferroviário na Matriz de Transporte de um país. Por sua vez, o planejamento operacional refere-se às ações adotadas para alcançar objetivos de curto e curtíssimo prazo. Um exemplo de objetivo operacional seria a determinação da frequência de ônibus em uma linha do transporte público para atender a quantidade de passageiros da hora-de-pico. Por último, o planejamento tático se preocupa com objetivos de médio prazo, funciona como um intermediário entre o nível operacional e o nível estratégico.

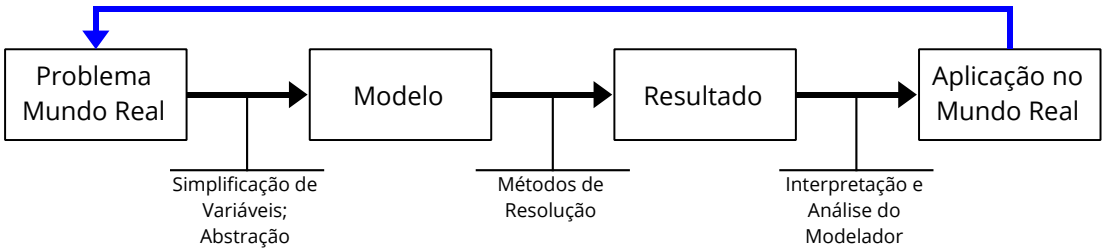
Um dos maiores desafios em relação ao planejamento de sistemas de transportes, independente do horizonte temporal, é a caracterização da demanda atual e futura do sistema em questão. Isto porque, primeiro, é baseado na demanda que as questões de dimensionamento do sistema (oferta) são respondidas e, segundo, a demanda por transportes é variável. No nível mais desagregado possível de informação, o individual, pessoas apresentam necessidades diferentes em relação às suas viagens. Como por exemplo, o período do dia em que se deslocam, o dia da semana, o tipo de modal que utilizam e o itinerário realizado (ORTÚZAR; WILLUMSEN, 2011).

Sendo assim, a caracterização da demanda por transportes, ou de forma mais simples, a caracterização de uma viagem, deve ser concebida além de um simples deslocamento espacial entre dois pontos, passando a ser dependente de diferentes variáveis, tais como: motivo, hora de partida, modal de transporte e itinerário. Esta característica multidimensional da demanda resulta, no âmbito pessoal, na escolha de um serviço de transporte que melhor atenda às necessidades da viagem; e, no âmbito do planejamento, torna a caracterização e a previsão da demanda uma tarefa desafiadora (CAMPOS, 2013).

Na busca por representar fenômenos do mundo real impossíveis de serem tratados analítica ou matematicamente em sua infinidade de variáveis, pesquisadores recorrem ao processo de modelagem. O processo de modelagem consiste em criar uma versão simplificada (um modelo) de um fenômeno do mundo real, levando em consideração somente as caracteríscas relevantes em relação a determinado aspecto do fenômeno, Figura 1. Desta forma, um problema complexo é transformado em um problema mais simples, passível de solução matemática e (ou) computacional, para que então, a solução encontrada para o modelo seja transferida ao problema real (HILLIER; LIEBERMAN, 2013).

A construção de um modelo é sempre dependente do ponto de vista que se pretende analisar o problema. Por exemplo, um garfo e uma faca sobre uma mesa poderiam ser utilizados como um modelo físico para representar a *trajetória* de dois veículos antes e após uma colisão. Todavia, os mesmos objetos não poderiam ser

Figura 1 – Processo de Modelagem



Fonte: Adaptado de Hillier e Lieberman (2013).

utilizados caso o propósito fosse determinar o *coeficiente de arrasto* do ar na superfície dos veículos (ORTÚZAR; WILLUMSEN, 2011).

Grande parte do processo de modelagem se concentra na determinação de um objetivo claro de análise (no exemplo anterior, a trajetória de dois veículos antes e depois de uma colisão), na escolha do tipo de modelo, como por exemplo, físico, matemático ou computacional, e das variáveis relevantes para o problema (que no exemplo poderia ser o ângulo formado entre os dois veículos).

Ortúzar e Willumsen (2011) argumentam que, apesar de a modelagem ser uma ferramenta robusta na resolução de problemas complexos do mundo real, como por exemplo, na estimativa da demanda, o processo de modelar não deve ser confundido com planejamento. O objetivo do planejamento de transportes de curto, médio ou longo prazo é minimizar os efeitos negativos gerados pelas atividades associadas ao ramo, tais como, desperdício de recursos naturais, congestionamentos, poluição e acidentes (EUROPEAN COMISSION, 2013; DEPARTMENT OF INFRASTRUCTURE AND REGIONAL DEVELOPMENT, 2014; BRASIL, 2013).

O modelo não substitui o tomador de decisão. Nesse sentido, a modelagem é uma parte do planejamento e funciona como uma ferramenta no processo de tomar decisões mais eficientes. Por exemplo, um modelo seria capaz de caracterizar a demanda de transportes atual em uma região, bem como, fornecer uma previsão de seu estado futuro, servindo como um indicador da variação da demanda, o que por sua vez, poderia apontar a necessidade de mudança para um modal de maior capacidade, caso a variação fosse positiva.

O conhecimento da demanda por transportes em sua multidimensionalidade (variação horária, diferenciação modal e distribuição espacial) é um elemento fundamental do planejamento (ORTÚZAR; WILLUMSEN, 2011). Desta forma, com o intuito de viabilizar o estudo da demanda, diversos modelos têm sido propostos nesta área do conhecimento. O processo de modelagem de demanda mais difundido atualmente é o *Modelo Sequencial*, também denominado *Modelo das Quatro Etapas*, que será tratado na próxima seção (CAMPOS, 2013; ORTÚZAR; WILLUMSEN, 2011).

2.2 Modelo das Quatro Etapas aplicado ao Planejamento de Transportes

O modelo de caracterização e previsão de demanda de transportes mais difundido atualmente é denominado Modelo Sequencial, ou Modelo das Quatro Etapas. O Modelo se resume em dividir uma região geográfica em sub-regiões, nomeadas de zonas de tráfego, de forma que estas apresentem características socioeconômicas homogêneas, para que então cada uma das quatro etapas do modelo sejam aplicadas de forma sequencial, Figura 3.

Devido ao alto custo do levantamento de dados para se determinar a demanda por transporte no nível individual em uma região, torna-se necessário a aplicação de modelos de agregação, em que um grupo de indivíduos é representado por um indivíduo médio (ORTÚZAR; WILLUMSEN, 2011). Por exemplo, indivíduos residentes em um mesmo domicílio podem ser representados pelas características médias de uma unidade residencial; várias unidades residenciais podem ser representadas pela média de um setor censitário; diversos setores censitários podem ser representadas pela média de um cidade, e assim sucessivamente. Por isso, antes da aplicação das quatro etapas propriamente ditas, existe uma "etapa 0" denominada de etapa de agregação ou zoneamento.

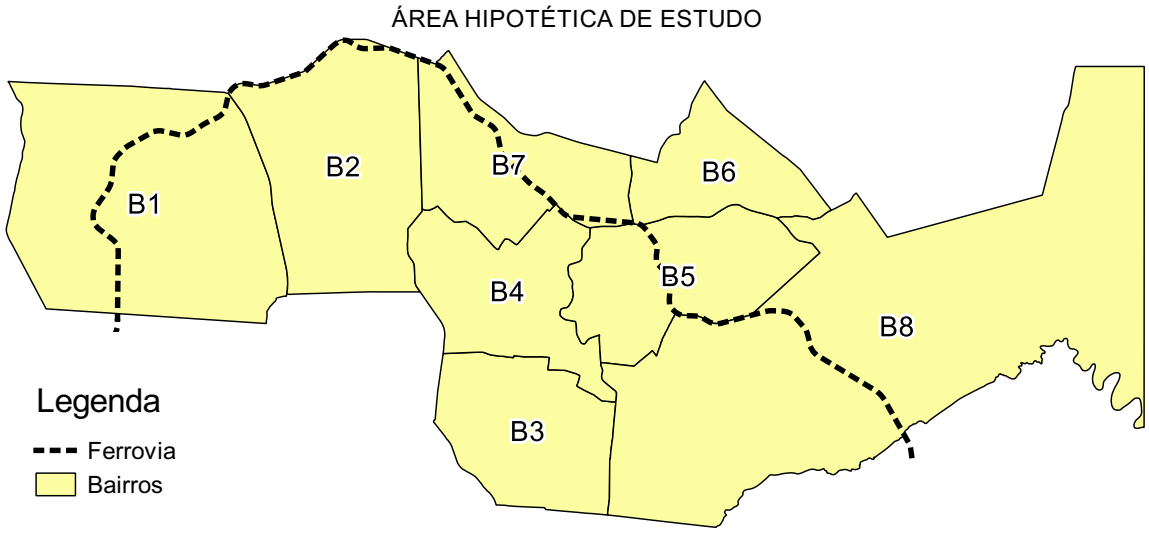
A etapa de agregação consiste em subdividir a área de estudo em n zonas, as quais possuirão um centróide, para que se identifique os aspectos socioeconômicos médios e de tráfego por unidade de tempo em cada uma das zonas. O centróide é um ponto que representa o centro geográfico ou, por exemplo, o ponto em que se concentra a maior parte dos deslocamentos da sub-região. Para efeitos de simplificação, as características socioeconômicas do centroide representam a zona em sua totalidade e todo o tráfego produzido ou atraído é concentrado no centroide, ou seja, o centroide é a única origem e o único destino das viagens de uma zona (CAMPOS, 2013).

As variáveis socioeconômicas são definidas como independentes; alguns exemplos são: população, renda média, número de domicílios, quantidade de empregos e quantidade de vagas em instituições de ensino. Por sua vez, as variáveis de tráfego são definidas como dependentes e são o total de viagens que saem e o total de viagens que chegam por unidade de tempo em cada zona de tráfego (CAMPOS, 2013).

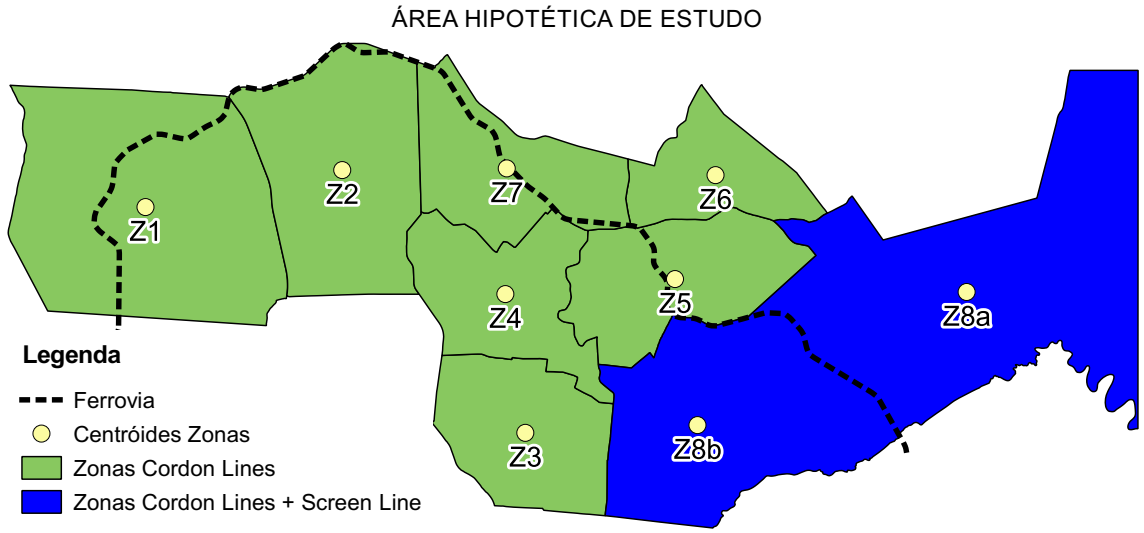
Para a realização deste zoneamento, Ortúzar e Willumsen (2011) sugerem a utilização de linhas virtuais, denominadas *cordon lines*, que podem ter sido estabelecidas em algum momento anterior ao estudo, como por exemplo os limites de um bairro ou distrito, ou através de barreiras físicas existentes denominadas *screen lines*, como por exemplo um corpo d'água ou uma linha férrea, ou ainda, ambos os tipos de linhas. A Figura 2 apresenta um exemplo de zoneamento para uma área de estudo hipotética utilizando o conceito de *cordon lines* e *screen lines*.

Como pode ser observado na Figura 2(a), a área de estudo hipotética já possuia

Figura 2 – Processo de zoneamento área de estudo hipotética.



(a) Área de estudo antes do zoneamento



(b)Área de estudo depois do zoneamento

Fonte: Autor, 2018.

divisões antes do zoneamento *(limites dos bairros)*. Após o zoneamento, Figura 2(b), as zonas preenchidas na cor verde mantiveram os limites virtuais pré-estabelecidos *cordon lines* e a zonas $Z8a$ e $Z8b$, preenchidas em azul, são proveniente da divisão do Bairro $B8$ em duas partes por uma linha férrea *screen lines*. Não existe uma regra para a delimitação do tamanho das zonas; a experiência do modelador e o conhecimento da região são importantes quesitos nessa fase. Entretanto, a partição adequada da área de estudo resultará em zonas com características socioecônomicas homegêneas em relação aos indivíduos cincunscritos dentro de cada zona (CAMPOS, 2013).

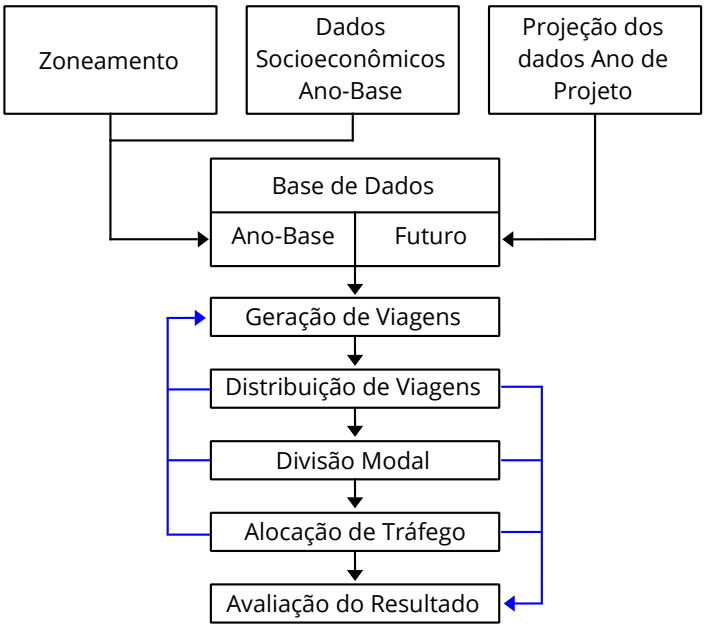
Além de tornar possível o levantamento de características socioeconômicas de

forma escalonada, o zoneamento da região de estudo possibilita averiguar, através, por exemplo, de contagens de tráfego junto aos limites das zonas, ou pesquisas domiciliares, a direção dos deslocamentos que cruzam estes limites. As viagens que ultrapassam os limites zonais em direção ao centroide são denominadas viagens atraídas e, as viagens que cruzam os limites para fora, são assumidas originárias no centróide e são denominadas viagens produzidas (ORTÚZAR; WILLUMSEN, 2011).

Após realizada a etapa de zoneamento e levantamento de dados, dá-se início a aplicação do Modelo Sequencial propriamente dito. O termo sequencial deriva do fato de que o resultado de uma etapa é o ponto de partida da outra, Figura 3 (CAMPOS, 2013). As etapas compreendidas no Modelo, de forma ordenada, são as seguintes:

1. Geração de Viagens: etapa que estima a quantidade total de viagens produzidas e atraídas em cada zona em função das características socioeconômicas.
2. Distribuição de Viagens: possui por objetivo indicar a quantidade de viagens entre pares de zonas.
3. Divisão Modal: estipula a porcentagem das viagens que são realizadas por um dado tipo de modal.
4. Alocação de viagens: caracteriza a trajetória das viagens realizadas por automóveis em uma rede de transporte.

Figura 3 – Modelo das Quatro Etapas



Fonte: Adaptado de Ortúzar e Willumsen (2011).

2.2.1 Etapa de Geração de Viagens

A primeira etapa do Modelo Sequencial é a etapa de Geração de Viagens. “A importância desta etapa está no fato de que seus resultados são o ponto de partida de todo o modelo sequencial, assim, deve-se cuidar para que o resultado desta etapa seja o mais preciso possível” (CAMPOS, 2013, p. 50).

Considerando que uma região foi dividida em n zonas e que para cada uma destas foram determinadas as t características socioeconômicas e o volume V_i de viagens produzidas e atraídas em cada centróide, a etapa de Geração de Viagens possui por objetivo construir um modelo matemático que seja capaz de relacionar as t variáveis socioeconômicas com a quantidade de viagens V_i geradas pela zona i (ORTÚZAR; WILLUMSEN, 2011).

Campos (2013) aponta que os objetivos de um modelo de Geração de Viagens são, de forma geral:

- Estimar a quantidade de viagens produzidas e atraídas em uma zona de tráfego com base nas variáveis socioeconômicas;
- Extrapolar o modelo para regiões diferentes da área de estudo que possuem características socioeconômicas semelhantes;
- Realizar previsões para o ano de projeto das viagens produzidas e atraídas através da extrapolação das variáveis socioeconômicas para o futuro.

Uma forma de relacionar as variáveis socioeconômicas com as de tráfego é construir uma equação linear, através do método de Regressão Linear Múltipla, Equação (2.1) (ORTÚZAR; WILLUMSEN, 2011).

$$V_i = \sum_s a_s x_{si} + b \qquad \forall i = 1, 2..., n \qquad \forall s = 1, 2..., t. \tag{2.1}$$

Onde:

V_i - Número de Viagens na zona i;
a_s - Coeficientes angulares da s-ésima variável socioeconômica;
x_{si} - Valor da s-ésima variável socioconômica da zona i;
b - Coeficiente linear.

Os coeficientes angulares e lineares são encontrados por calibração utilizando os dados do ano-base de todas as zonas de tráfego (CAMPOS, 2013).

A principal vantagem de um modelo de Geração de Viagens, construído através de regressão linear múltipla, é a sua dependência em relação às variáveis socioeconômicas. Isto implica que extrapolando as variáveis explicativas para um período futuro é possível estimar a variação do número de viagens. Outra consequência da dependência entre a geração de viagens e as variáveis socioeconômicas é que

o modelo construído em uma área de estudo pode ser aplicado em outra região que possua características socioeconômicas semelhantes sem a necessidade de novas contagens de tráfego para a calibração dos coeficientes angulares e lineares (ORTÚZAR; WILLUMSEN, 2011).

2.2.2 Etapa de Distribuição de Viagens

A segunda etapa do Modelo Sequencial refere-se à etapa de Distribuição de Viagens. O objetivo desta fase é estimar o número de viagens entre pares de zonas de tráfego V_{ij} que possuem como origem a zona i e destino a zona j para cada par ij de zonas, com $ij = 11, 12..., 1n, ..., n1, ..., nn$, de forma que:

$$O_i = V_{i1} + V_{i2} + ... + V_{in} = \sum_j V_{ij} \qquad \forall\, i = 1, 2, ..., n. \tag{2.2a}$$

$$D_j = V_{1j} + V_{2j} + ... + V_{nj} = \sum_i V_{ij} \qquad \forall\, j = 1, 2, ..., n. \tag{2.2b}$$

Onde:

O_i - Número de Viagens com origem na zona i;
D_j - Número de Viagens com destino na zona j;
V_{ij} - Número viagens entre pares de zonas de tráfego.

Das Equações (2.2a) e (2.2b) pode ser derivada a matriz denominada Matriz Origem-Destino, ou Matriz O/D, conforme Tabela 1. Existem vários métodos para se determinar os elementos V_{ij} de uma matriz O/D, para o ano de projeto. Para o caso no qual a Matriz O/D se encontra determinada para o ano-base os métodos mais utilizados são os Métodos de Fatores de Crescimento Uniforme e Médio, e Método de Fratar (CAMPOS, 2013). Em oposição, para o caso em que a Matriz O/D precisa ser caracterizada também para o ano-base, pode-se utilizar Modelos Gravitacionais de Distribuição de Viagens (ORTÚZAR; WILLUMSEN, 2011; CASEY, 1955) que é baseado na Lei gravitacional de Newton.

Tabela 1 – Matriz Origem Destino genérica

	1	2	3	$\dots$	j	$\dots$	n	$\sum_j V_{ij}$
1	V_{11}	V_{12}	V_{13}	$\dots$	V_{1j}	$\dots$	V_{1n}	O_1
2	V_{21}	V_{22}	V_{23}	$\dots$	V_{2j}	$\dots$	V_{2n}	O_2
3	V_{31}	V_{32}	V_{33}	$\dots$	V_{3j}	$\dots$	V_{3n}	O_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
i	V_{i1}	V_{i2}	V_{i3}	$\dots$	V_{ij}	$\dots$	V_{in}	O_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
n	V_{n1}	V_{n2}	V_{n3}	$\dots$	V_{nj}	$\dots$	V_{nn}	O_n
$\sum_i V_{ij}$	D_1	D_2	D_3	$\dots$	D_j	$\dots$	D_n	$\sum_{ij} V_{ij}$

Fonte: adaptado de Ortúzar e Willumsen (2011).

Para o caso de distribuição de viagens, a “força de atração” entre dois centroides seria igual à quantidade de viagens V_{ij}; a “massa” representa aspectos socioeconômicos e de tráfego de cada uma das zonas e o “quadrado da distância” representa o custo generalizado de transporte entre dois centróides (CAMPOS, 2013).

2.2.3 Etapa de Divisão Modal

Considerando-se que existem m modos de transportes disponíveis na região de projeto, a etapa de Divisão Modal objetiva estimar a proporção das viagens V_{ij} entre pares de zonas que utilizam o modal m, o que resulta em V_{ij}^k, Equação (2.3) (ORTÚZAR; WILLUMSEN, 2011). Sendo assim, após a etapa de Divisão Modal é possível derivar matrizes O/D por tipo de modal para a área de projeto.

$$V_{ij} = V_{ij}^1 + V_{ij}^2 + ... + V_{ij}^k + ... + V_{ij}^m = \sum_k V_{ij}^k \qquad \forall\, i, j \qquad \forall\, k = 1, 2..., m. \tag{2.3}$$

Onde:

V_{ij} - Número de Viagens com origem na zona i e destino na zona j;

V_{ij}^k - Número de Viagens com origem na zona i e destino na zona j que utiliza o modal k.

Campos (2013) afirma que existem basicamente dois tipos de métodos para a determinação da divisão modal: determinísticos e probabilísticos. Os métodos determinísticos, de forma geral, se utilizam de regressão para calcular a proporção de viagens realizadas por cada modo de transporte, relacionando a divisão modal com

características socioeconômicas, como por exemplo, o grau de motorização de uma região, como pode ser observado na Equação (2.4) apresentada por Campos (2013).

$$y_i = 0,877 - 0,00086x_i \tag{2.4}$$

Onde:

y_i - percentual de viagens por transporte coletivo na zona i;
x_i - grau de motorização da zona i (número de veículos por 1000 habitantes).

Por sua vez, os métodos probabilísticos relacionam a proporção de viagens realizadas através de um dado modal com a probabilidade de escolha do modal em questão. Por consequência, os modelos probabilísticos também são denominados modelos de escolha discreta que está fundamentado na relação entre a *utilidade* de dois ou mais modos de transporte (ORTÚZAR; WILLUMSEN, 2011).

A probabilidade da escolha de um meio de transporte k, dentre um total de m modais, é baseada na *Teoria de Utilidade* (ORTÚZAR; WILLUMSEN, 2011). Novaes (1986) define Função Utilidade, sob o ponto de vista do estudo de demanda, como uma expressão matemática que determina o grau de satisfação que o usuário de um dado modo transporte alcança com a escolha do mesmo. As funções utilidades são determinadas com base em dados levantados através de Pesquisas de Preferência Revelada, e Pesquisas de Preferência Declarada (ORTÚZAR; WILLUMSEN, 2011). Um exemplo de função utilidade pode ser encontrado em (2.5).

$$U_k = \alpha - at_k - bc_k \tag{2.5}$$

Onde:

U_k - Utilidade do modal k;
α - Utilidade intrínseca do modal k;
a - coeficiente relacionado ao Tempo de Viagem t_k;
t_k - Tempo de viagem do modal k;
b - coeficiente relacionado ao Custo de Viagem c_k;
c_k - Custo de viagem do modal k.

Pode-se observar da Equação (2.5) que o tempo e o custo de viagem influenciam de forma negativa na utilidade do modal: quanto maior o tempo e o custo, menor será a utilidade. Após a determinação das funções utilidades, dois tipos de modelos de escolha discreta podem ser aplicados, a saber, Logit ou Probit, que podem ser Binomiais, quando existem apenas duas alternativas de modos de transporte, ou Multinomiais, quando existem mais de dois modais. O modelo mais utilizado devido a sua simplicidade matemática é o Modelo Logit, Equação (2.6) (ORTÚZAR;

WILLUMSEN, 2011).

$$P_{kq} = \frac{e^{U_{kq}}}{\sum_i e^{U_{iq}}} \qquad com \quad i = 1, 2, ..., k, ..., m. \tag{2.6}$$

Onde:

P_{kq} - Probabilidade de escolha do modal k pelo indivíduo q;
U_{kq} - Utilidade do modal k para o indivíduo q;
U_{iq} - Utilidades dos modais i para o indivíduo q.

Da Equação (2.6) pode-se observar que a probabilidade de escolha de um modal k é calculada como uma relação entre sua utilidade e a dos demais modos de transporte disponíveis. Multiplicando a Matriz O/D encontrada na etapa de Distribuição de Viagens por todos os P_{kq} encontrados na etapa de Divisão Modal, obtêm-se uma matriz O/D por tipo de modal.

2.2.4 Etapa de Alocação de Tráfego

A última etapa do Modelo Sequencial é denominada Alocação de Tráfego. Ortúzar e Willumsen (2011) expõem que o objetivo desta fase é avaliar a distribuição do fluxo de viagens em uma rede. Considerando-se que entre um par de zonas de tráfego ij existem u rotas disponíveis, a fase de Alocação de Tráfego determinará a proporção das viagens realizadas pelo modo k entre o par O/D ij (V_{ij}^k) que utiliza cada rota r, o que resulta em V_{ij}^{kr} (Equação (2.7)). Sendo assim, o problema de Alocação se caracteriza por um problema de escolhas de rotas entre pares O/D e a distribuição do fluxo V_{ij}^k entre as r rotas disponíveis Campos (2013)

$$V_{ij}^k = V_{ij}^{k1} + V_{ij}^{k2} + ... + V_{ij}^{ku} = \sum_r V_{ij}^{kr} \qquad \forall\ k \qquad com \quad r = 1, 2, ..., u. \tag{2.7}$$

Onde:

V_{ij}^{kr} - Número de Viagens com origem na zona i e destino na zona j que utiliza o modal k e a rota r.

Campos (2013) afirma que a fase de Alocação pode ser realizada para todos os k modos de transporte que estiverem disponíveis na área de projeto. Contudo, os modelos clássicos de alocação têm como foco principal a alocação de automóveis em vias urbanas, devido ao fato de que a maior parte das viagens nos centros urbanos são realizadas através deste modal.

Segundo Sheffi (1985), o problema de alocação consiste basicamente na resolução de modelos de minimização não lineares com restrições de igualdade e não-

negatividade que possuem como base os critérios de alocação equilibrada definidos por Wardrop (1952), que são:

- Os tempos de viagens nas rotas não utilizadas entre um par O/D são sempre maiores ou iguais aos tempos de viagens nas rotas utilizadas;
- Os tempos de vigens de todas as rotas utilizadas entre um par O/D são iguais e mínimos sob dado carregamento de rede.

2.2.5 Considerações Gerais a respeito do Modelo das Quatro Etapas

Através da Equação (2.7) é possível perceber que ao final da aplicação do Modelo das Quatro etapas têm-se a demanda de transporte estimada em sua multimensionalinidade; uma viagem é definada em termos dos locais de origem e destino, modo de transporte utilizado e itinerário.

Campos (2013) e Ortúzar e Willumsen (2011) compartilham do ponto de vista que a aplicação do Modelo Sequencial em sua forma *clássica*, isto é, todas as etapas, Geração, Distribuição, Divisão e Alocação, aplicadas de forma ordenada, possui maior relevância no planejamento estratégico dos transportes, do que no planejamento operacional dos modais presentes na área em questão. Primeiramente, devido ao alto custo para aquisição de dados (contagens de tráfego, pesquisas domiciliares, por exemplo) o Modelo das Quatro etapas é comumente aplicado ao período de pico, o que, do ponto de vista operacional, ocasionaria o super dimensionamento da demanda para as demais horas do dia, resultando em desperdício de recursos.

Segundo, em sua forma clássica, o Modelo Sequencial possui alto grau de abrangência, pois explica a interação entre três fatores: características socioeconômicas, diferentes modais de transporte e a rede viária de uma região. Por último, a variação da demanda para longos períodos de tempo é menor do que a variação para períodos curtos. Por exemplo, a variação do número médio de passageiros de ônibus, na hora-pico, de uma linha de transporte público é menor se comparada com a média para o mesmo horário no ano seguinte, do que se comparada com a média de passageiros da hora subsequente à hora-pico. Sendo assim, o Modelo das Quatro Etapas, em sua forma clássica, funciona de forma mais eficiente como um auxiliador nas decisões de médio e longo prazo, como por exemplo, no direcionamento dos investimentos de recursos, mudanças na infraestrutura viária de uma região e priorização de determinados modais de transporte em detrimento de outros.

Ortúzar e Willumsen (2011) argumentam que o Modelo das Quatro etapas não é um modelo matemático em si, mas sim uma forma de abordagem sequencial para o problema do estudo da demanda de transportes. Por sua vez, cada uma das etapas do Modelo pode ser resolvida por um ou mais modelos matemáticos. Em outras palavras, o Modelo das Quatros etapas é uma *abordagem* de resolução do problema de estudo de demanda que compreende diferentes modelos matemáticos como etapas. Por ser

um conceito, o Modelo das Quatro Etapas pode ser generalizado para outras formas de estudo da demanda. Por exemplo, caso se queira analisar a quantidade e rotas de viagens realizadas através de bicicletas entre duas zonas pré-determinadas, não são necessárias a modelagem da etapa de Divisão Modal e Distribuição de Viagens, o que resultaria em um modelo de Geração-Alocação.

Dado que o objetivo deste trabalho é utilizar um modelo de previsão de demanda que auxilie na *operação* do sistema de transporte público por ônibus, fica evidente que o Modelo Sequencial, em sua forma clássica, não é o modelo mais indicado para esta tarefa. Considerando-se que a origem, o itinerário e o destino das linhas de ônibus são fixos, o modelo que será utilizado pode ser classificado como um modelo de *Produção de Viagens* na origem das linhas, visto que, para efeitos de simplificação, será assumido que os usuários embarcam na origem de determinada linha e desembarcam somente no destino.

Para realizar a Previsão de Viagens (ou número de passageiros) na origem de linhas do transporte coletivo da Cidade de Joinville, Redes Neurais Artificiais, doravantes denominadas RNAs, serão utilizadas. Contudo, antes da contextualização a respeito de Redes Neurais faz-se necessário fundamentar a importância do conhecimento da demanda, assim como, do equilibrío entre esta e a oferta, para o transporte público. A próxima seção tratará sobre a relevância do transporte coletivo para os centros urbanos.

2.3 Transporte público coletivo: relação entre oferta e demanda de Viagens

Ferraz e Torres (2004) argumentam que a oferta de transporte é identificada como um serviço, razão pela qual não pode ser estocada para uso em momentos de alta demanda. Ortúzar e Willumsen (2011) compartilham da mesma ideia afirmando que o serviço de transporte deve ser consumido no momento e lugar em que é produzido, do contrário, o seu benefício é perdido. Por esta razão, quanto mais acurada e precisa for a estimativa da demanda, menor será o desperdício de recursos dos sistemas de transportes como um todo.

Segundo Ferraz e Torres (2004), as atividades relacionadas à educação, trabalho, saúde e lazer somente são possíveis através da movimentação de pessoas e bens. Assim, o transporte é tão importante como os serviços de abastecimento de água, tratamento de esgosto e distribuição de energia elétrica. Vuchic (2005) acrescenta afirmando que proporcionar adequada mobilidade a todos os cidadãos é um fator vital para o desenvolvimento econômico e social de uma cidade. Neste sentido, o serviço de transporte coletivo público tem como função primária ser um promotor de desenvolvimento, igualdade e inclusão social, possibilitando o acesso às infraestruturas e serviços urbanos, a todos os cidadãos.

Devido a falta de uma base de dados centralizada de informações referentes ao transporte de passageiros dentro dos centros urbanos, é difícil afirmar com precisão a porcentagem de viagens realizadas por modos públicos de transporte dentro das cidades do território nacional. Contudo, observa-se que grande parte dos deslocamentos realizados dentro dos centros urbanos ocorrem através do transporte coletivo. Por exemplo, São Paulo (2018) mostra que aproximadamente 32% das viagens realizadas dentro da Cidade de São Paulo, que é a maior cidade brasileira com 12.106.920 de habitantes no ano de 2017 (IBGE, 2018b), são realizadas por modos públicos de transporte. Já para o município do Rio de Janeiro, a segunda maior cidade brasileira com 6.520.266 de habitantes no ano de 2017 (IBGE, 2018c), as viagens por modos público de transporte representam 51,1% de todas as viagens realizadas (RIO DE JANEIRO, 2013).

Em relação à exploração econômica do serviço de transporte público coletivo dentro das cidades brasileiras, a Lei Federal 12.578 de 2012 institui que a realização dessa modalidade de serviço deverá acontecer sob a forma de concessão ou permissão, através de processo licitatório realizado pelo orgão público competente e que o preço da tarifa a ser pago pelos usuários deve cobrir os reais custos do serviço prestado pelo operador público ou privado, além da remuneração, quando houver, do prestador (BRASIL, 2012).

Outro aspecto importante da Lei Federal 12.578 de 2012 é que a mesma institui o governo municipal como o agente responsável pela fixação de metas de qualidade (nível de serviço) e seus instrumentos de controle e avaliação visando gerenciar o conflito de interesses entre o usuário e a prestadora de serviço (BRASIL, 2012). À medida que o usuário busca minimizar os custos financeiros e temporais com o transporte e espera receber alto nível de serviço, a prestadora visa a maximização da receita obtida com a prestação do serviço. Por isso, é de extrema importância que o município determine um nível de serviço que, concomitantemente, incentive a utilização do transporte por parte dos usuário e permita a exploração econômica do negócio.

Enquanto a tarifa corresponde diretamente ao valor monetário pago para a utilização do sistema de transporte, o nível de serviço oferecido ao usuário é dependente de mais de uma variável. De acordo com Ferraz e Torres (2004), são doze as principais variáveis que influenciam no nível de serviço do transporte público urbano: acessibilidade, frequência de atendimento, tempo de viagem, lotação, confiabilidade, segurança, características dos veículos, características dos locais de parada, sistema de informações, conectividade, comportamento dos operadores e condição das vias.

Alguns dos fatores que determinam o nível de serviço global do transporte público elencados acima são subjetivos e, por consequência, difíceis de serem avaliados metricamente de forma direta. Como é o caso, por exemplo, das características dos locais de paradas, segurança e comportamento dos operadores. Entretanto, outros

fatores são passíveis de medição e fornecem uma boa indicação da qualidade do sistema de tranporte, como por exemplo, tempo de viagem (intervalo de tempo entre a origem e o destino da viagem), freqüência de atendimento (número de vezes que um ônibus atravessa determinada seção do itinerário por unidade de tempo), e lotação (quantidade de passageiros por unidade de área do veículo).

Para Ferraz e Torres (2004) a qualidade do transporte público urbano e a sua sustentabilidade se encontra no equilíbrio entre o nível de satisfação de todos os atores direta ou indiretamente envolvidos no sistema: usuários, comunidade, governo, trabalhadores do setor e prestadoras de serviço. A insatisfação de um dos grupos leva, inevitavelmente, ao desequilíbrio do sistema; o que pode resultar na queda da demanda, aumento do preço da tarifa, aumento da lotação e greves.

Levando-se em consideração os dois atores principais do serviço de transporte público, o usuário e a prestadora de serviço, e um dado nível de serviço estipulado pelo orgão governamental competente, o adequado atendimento da demanda de passageiros (lotação) é um dos principais meios que a operadora possui de, simultaneamente, garantir a manutenção do nível de serviço e realizar sua atividade comercial de forma eficiente e eficaz (VUCHIC, 2005).

A eficiência na produção de um bem ou serviço diz respeito à produtividade expressa, de maneira geral, pela relação entre a quantidade de unidades produzidas e os insumos gastos na produção (EVANS; LINDSAY, 2014). No caso do serviço de transporte público, as unidades produzidas são as viagens ofertadas e os insumos são por exemplo os veículos, combústivel e mão de obra necessária para a realização dessas viagens. Sob a luz do conceito de eficiência econômica, é trivial considerar que, dado um nível de serviço com uma taxa de lotação *constante*, quanto mais próximo for a quantidade de passageiros do limite de lotação, mais eficiente será o sistema, já que os custos fixos para se realizar uma viagem seriam diluídos por uma quantidade maior de usuários.

Apesar de matematicamente um número de usuários maior que o limite de lotação estabelecido pelo nível de serviço resultar em uma eficiência econômica maior para uma dada viagem, na prática, observa-se que esta situação acarretá na insastisfação por parte dos usuários devido ao desconforto decorrente da excessiva proximidade entre as pessoas, o que a médio e longo prazo pode refletir na mudança da alternativa de transporte por parte do usuário. Sendo assim, o adequado conhecimento da demanda (previsão), bem como o atedimento da mesma, constitui extrema importância para a sustentabilidade do sistema de transporte público de uma cidade (FERRAZ; TORRES, 2004).

O ônibus é o modal mais utilizado no Brasil, aproximadamente 59% das cidades brasileiras possuem o ônibus como meio de transporte coletivo público (NTU, 2017). Ao longo dos últimos anos, Sistemas Inteligentes de Transporte têm sido

incorporados a esta modalidade com o intuito de torná-la mais eficiente. Como é o caso da bilhetagem eletrônica, ou *smart card*, que pode ser classificado como um *Sistema Avançado de Transporte Público (APTS)*, que têm por objetivo automatizar o processo de gerenciamento da tarifa, conferindo ao sistema redução de custo com mão-de-obra, diminuição dos tempos de embarque e controle sobre a geração de receita. Atualmente, a bilhetagem eletrônica está presente em 86,5% das cidades que possuem o ônibus como alternativa de transporte público (NTU, 2017).

Além de tornar o transporte por ônibus mais eficiente, a bilhetagem eletrônica gera grandes volumes de dados de viagens. Para cada viagem registrada através do *smart card* têm-se associados atributos espaço temporais como, por exemplo, a origem da viagem e o tempo inicial de sua realização. Estes dados podem ser utilizados na construção de modelos de previsão de demanda de curto prazo; previsão diária ou horária, sendo um auxiliador na operação do sistema, especificamente, na determinação da oferta de assentos (TSUNG-HSIEN; CHI-KANG et al., 2009; DRAGANA; MILICA et al., 2015).

Devido ao grande volume de dados gerados pelos APTS, métodos convencionais de análise mostram-se pouco eficientes no tratamento destes dados, sendo necessária a aplicação de métodos de análise da área de Mineração de Dados (TAN; STEINBACH; KUMAR, 2005). A próxima seção objetiva descrever como modelos de RNAs têm sido utilizadas para a construção de modelos de previsão de curto prazo.

2.4 Mineração de Dados Aplicada ao Planejamento Operacional de Transportes

De acordo com Tan, Steinbach e Kumar (2005) a mineração de dados é uma área do conhecimento que combina métodos tradicionais de análises de dados provenientes da estatística, tais como, amostragem, estimação e teste de hipóteses, com algoritmos sofisticados das áreas de inteligência artifical e aprendizado de máquina, para o processamento de grandes volumes de dados com o intuito de encontrar informações que permaneceriam desconhecidas após a aplicação de métodos tradicionais.

Além da capacidade de tratar volumes de dados maiores do que aqueles que seriam possíveis através de métodos usuais, Tan, Steinbach e Kumar (2005) apresentam outras duas vantagens que a mineração de dados apresenta em relação aos métodos tradicionais. Primeiramente, as técnicas de mineração de dados têm a capacidade de analisar dados que possuem alto grau de dimensão e heterogeneidade (TAN; STEINBACH; KUMAR, 2005). Isto é, cada registro (uma unidade de dado) pode conter centenas de atributos (alta dimensionalidade), de diferentes tipos (heterogeneidade), como por exemplo, discretos ou contínuos. Segundo, as técnicas de mineração de dados permitem a análise de diversas hipóteses simultaneamente. O

método estatístico tradicional é baseado no paradigma de *hipótese-teste* em que uma hipótese é proposta, um experimento é projetado para aquisição de dados e então os dados são testados em relação à hipótese inicial. Caso seja necessário a avaliação de uma nova hipótese, provavelmente, um novo experimento deverá ser projetado para nova aquisições de dados, o que torna o processo tradicional de teste de hipóteses dispendioso.

Tan, Steinbach e Kumar (2005) afirmam que as tarefas desempenhadas pelas ferramentas de mineração podem ser divididas, de forma geral, em duas categorias: predição e descrição. As tarefas de descrição objetivam delinear padrões de comportamento dos dados: tendências, correlações, agrupamentos e anomalias. Por sua vez, as análises de predição têm por finalidade estimar o valor de uma variável *alvo* (ou dependente) através de um modelo construído com base nos valores de outras variáveis (ou atributos), denominadas explicatórias ou independentes.

Em se tratando de aprendizado supervisionado, no qual existem observações que podem ser utilizadas para especificar o comportamento desejado de um modelo, existem duas principais categorias de modelos. Quando a variável alvo for discreta e possuir um conjunto finito e predefinido de valores (classes), tem-se um modelo de classificação. Nos casos em que as variáveis podem assumir um conjunto de valores não previamente determinados (não há uma noção de classes), tem-se um modelo de regressão. O objetivo de ambos os tipos de modelos é minimizar o erro entre o valor predito e o valor real da variável alvo (TAN; STEINBACH; KUMAR, 2005). Neste trabalho as RNAs serão utilizadas como modelos de regressão.

2.4.1 Redes Neurais Artificiais: Conceitos Básicos

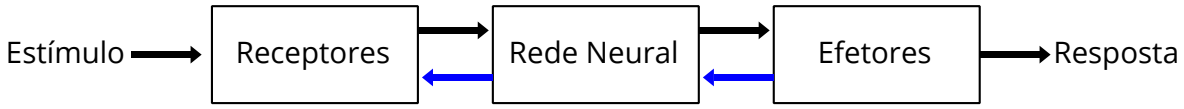
Haykin (2009) apresenta que o desenvolvimento de RNAs têm sido motivado pelo reconhecimento de que o cérebro humano processa determinados tipos de informações de forma mais eficiente se comparado a um computador digital. O cérebro humano é complexo, não-linear, possui capacidade de aprendizado e analisa dados de forma paralela (várias informações podem ser processadas simultaneamente), o que resulta, por exemplo, na capacidade que o ser humano possui de reconhecer padrões, generalizar o conhecimento e tomar decisões em resposta aos estímulos externos em menor tempo do que um computador. Sendo assim, as RNAs têm por objetivo a construção de modelo que *“imite”* a forma de processamento cerebral.

A capacidade de processamento e aprendizagem do cérebro humano deriva de células nervosas denominadas neurônios e na forma como estas células interagem entre si formando uma rede, denominada rede neural biológica. Esta rede é capaz de receber um estímulo de outra parte do ambiente externo, converter este estímulo em impulsos elétricos, que por sua vez serão convertidos em uma resposta ao estímulo. Após executada a resposta, a rede cerebral analisa os resultados obtidos, e aprende

com a experiência (HAYKIN, 2009).

A Figura 4 apresentada por Arbib (1987) ilustra o conceito de processamento de informação e aprendizagem cerebral discutida no parágrafo anterior. Na Figura 4 é possível observar que a propagação do sinal ocorre da esquerda para a direita (setas pretas) com o intuito de gerar uma resposta em relação a um estímulo. Por sua vez, a retropropagação do aprendizado ocorre no sentido contrário (setas azuis) com o objetivo de facilitar transmissão futura de sinais gerados pelo mesmo tipo de estímulo.

Figura 4 – Representação esquemática do sistema nervoso humano



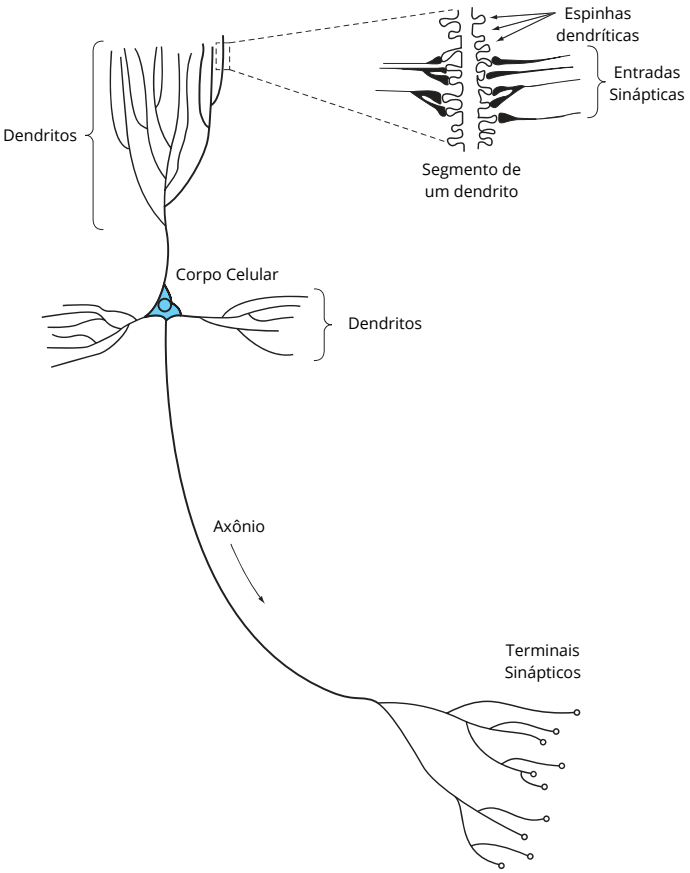
Fonte: Arbib (1987).

Os neurônios são as unidades fundamentais de processamento de uma rede neural, de forma que, para entender o funcionamento de uma rede formada por este tipo de célula, é necessário primeiramente entender como funciona, de forma muito simplificada, uma única unidade de processamento. Um esquema simplificado de uma célula nervosa é apresentado na Figura 5.

Na Figura 5 observa-se que os neurônios podem ser divididos anatomicamente, de modo simplificado, em quatro partes: dendritos, corpo da célula, axônio e terminações sinápticas. O processamento de informação têm início nos dendritos e fim nas terminações sinápticas. Biologicamente, cada neurônio possui a capacidade de funcionar como um receptor e transmissor de informação. O fluxo de informação ocorre da seguinte forma: os dendritos são responsáveis por receber as entradas sinápticas (estímulos) que ocorrem através da passagem de substâncias químicas, provenientes do ambiente externo à celula ou de terminações sinápticas de outros neurônios, pelas espinhas dendríticas; um neurônio cerebral pode receber aproximadamente 10.000 conexões em seus dendritos. As substâncias químicas são transportadas para o corpo da célula onde são convertidas em pulsos elétricos, denominados *potenciais de ação*. Os pulsos elétricos são transportados pelo axônio até as terminações sinápticas. As terminações sinápticas são responsáveis por converter o potencial de ação em substâncias químicas novamente e propagar a informação para os neurônios subsequentes (HAYKIN, 2009).

A Figura 6 apresenta um *modelo* artificial de neurônio (HAYKIN, 2009). Três elementos básicos podem ser identificados na Figura 6 que representa um neurônio artificial denotado pelo índice k: um grupo de m sinapses ou conexões, uma estrutura de somatório e uma função de ativação $\varphi(.)$. O grupo de conexões corresponde aos sinais x_j proveniente da $j-\text{ésima}$ conexão com o neurônio k, com $j = 1, 2, ..., m$.

Figura 5 – Anatômia Simplificada de um Neurônio Humano



Fonte: Haykin (2009).

É importante observar que os sinais x_j são multiplicados por pesos sinápticos w_{kj}; o primeiro índice em w_{kj} corresponde ao neurônio em questão e o segundo indíce aponta a conexão propriamente dita. Os pesos servem para permitir que determinadas conexões sejam mais ou menos importantes que outras, o que é fundamental para o aprendizado da rede (HAYKIN, 2009).

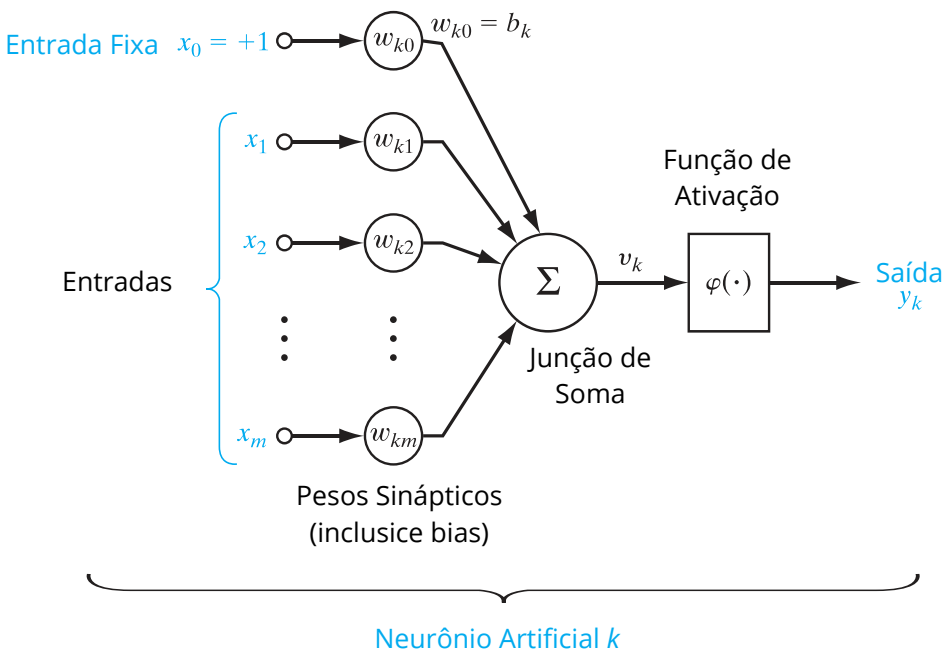
A estrutura de somatório da Figura 6 funciona como um *combinador linear* que soma todas as entradas $x_j w_{kj}$ de modo a produzir um potencial de ativação v_k (Equação 2.8), que será utilizado pela função de ativação $\varphi(.)$ para produzir a saída y_k (Equação 2.9) (HAYKIN, 2009).

$$v_k = \sum_j w_{kj} x_j \qquad \forall\ j \tag{2.8}$$

Onde:

v_k - potencial de ativação do neurônio k;
w_{kj} - peso da sinapse j do neurônio k;
x_j - valor do sinal proveniente da sinapse j.

Figura 6 – Modelo Artificial de Neurônio



Fonte: Haykin (2009).

$$y_k = \varphi(v_k) \tag{2.9}$$

Onde:

y_k - saída do neurônio k;
$\varphi(.)$ - função de ativação do neurônio k.

O objetivo da função de ativação é limitar a amplitude da saída y_k de um neurônio k no intervalo fechado [0,1]. Dois tipos básicos de função de ativação são apresentadas por Haykin (2009). Função Degrau, Equação (2.10a) e Figura 7(a), e Função Sigmoidal. Um exemplo de Função Sigmóide é a Função Logística, Equação (2.10b) e Figura 7(b).

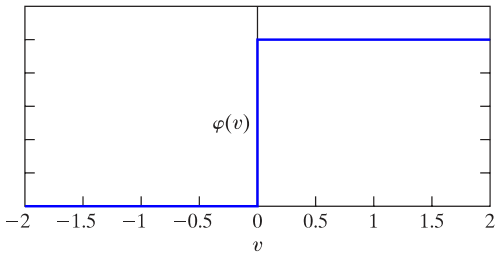
$$y_k = \begin{cases} 1, \text{ se } v_k \geq 0 \\ 0, \text{ se } v_k < 0 \end{cases} \tag{2.10a}$$

$$y_k = \frac{1}{1 + e^{-av_k}} \tag{2.10b}$$

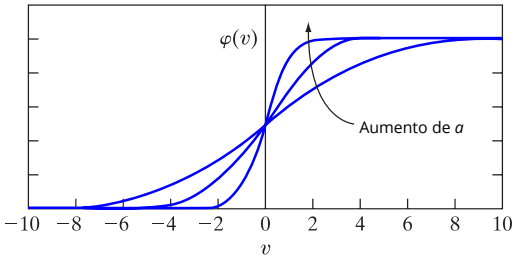
Onde:

$\varphi(v_k)$ - saída do neurônio k;

Figura 7 – Funções de Ativação



(a) Função Sinal



(b) Função Sigmoidal

Fonte: Haykin (2009).

v_k - potencial de ativação do neurônio k;
a - parâmetro de inclinação da Função Sigmoidal.

Da Equação (2.10a) pode-se observar que a Função Degrau restringe a saída y_k em 0 ou 1 baseado em um valor limiar de v_k (Figura 7). Haykin (2009) aponta que pelo fato dessa função não ser diferenciável a capacidade de aprendizado da rede neural fica comprometida, pois durante o processo de retropropagação do aprendizado o gradiente da função de ativação é utilizado para a calibração dos pesos w_{kj} das conexões. Como esta função não é contínua e, consequentemente, não diferenciável, não pode-se obter tal gradiente.

Por sua vez, a Função Sigmóide, que recebe esse nome devido ao formato de "S" de seu gráfico (Figura 7) possui ambas as características necessárias, sendo estas, não linearidade e diferenciabilidade, para tornar uma RNA robusta, tanto no sentido da complexidade do problema a ser resolvido como na capacidade de aprendizado (HAYKIN, 2009). Haykin (2009, p. 14) afirma que por estes motivos as funções do tipo sigmóide "são de longe o tipo mais comum de função de ativação utilizada na construção de redes neurais". É importante observar que a Função Logística também limita a saída y_k em 0 e 1 baseado no valor de v_k. Neste trabalho, a Função Sigmóide será usada como função de ativação dos neurônios artificiais.

Uma rede neural biológica é, de forma geral, o resultado da conexão de

células neurais; é estimado que o córtex humano possui aproximadamente 10 bilhões de neurônios que formam entre si 60 trilhões de conexões (SHEPHERD, 2004). Analogamente, uma RNA é o resultado da conexão de neurônios artificiais, que são definidos em função dos seus pesos e de sua função de ativação, Figura 6. A forma como essas conexões são realizadas determina o que é chamado de *Arquitetura* da rede neural. Uma forma comum de estruturar uma rede neural é através de camadas de neurônios (HAYKIN, 2009).

Na estruturação por camadas, a forma mais simples, seria uma rede neural formada por uma camada de entrada de sinal, que se liga diretamente a uma camada de saída, Figura 8(a). A este tipo de estrutura dá-se o nome de *Single-Layer Networks*, da qual a representante mais conhecida é a rede Perceptron de única camada, pois só existe uma camada que realiza processamento, sendo esta, a camada de saída. Caso a rede seja estruturada com mais de uma camada de processamento como na Figura 8(b), a mesma será classificada como Multilayer ou multicamada. Dentre as redes multicamada, a mais conhecida é a *Multilayer Perceptron (MLP)*, ou Perceptron de Múltipas Camadas (HAYKIN, 2009). As redes utilizadas no presente trabalho são do tipo MLP.

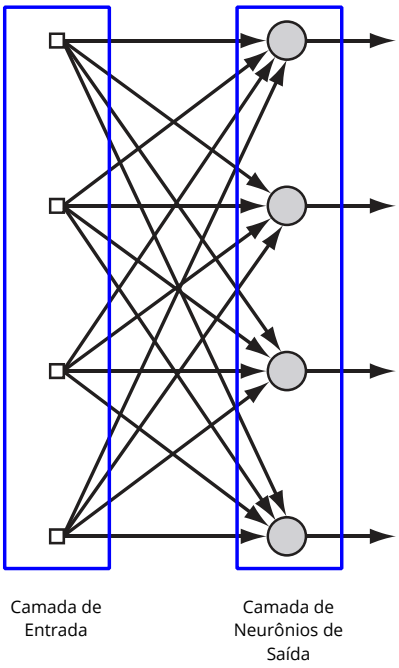
As camadas que estiverem localizadas entre a camada de entrada e a camada de saída serão denominadas camadas de neurônios escondidos. Uma rede com p nós de entrada, h_1 neurônios na primeira camada oculta, h_2 neurônios na segunda camada oculta (caso exista) e, q neurônios na camada de saída é denominada uma rede $p - h_1 - h_2 - q$ (HAYKIN, 2009).

Na representação gráfica de um RNA, Figura 8, as entradas são representadas por pequenos quadrados para ilustrar a ideia de que não é realizado processamento nesta camada, enquanto que os neurônios das camadas escondidas e da camada de saída são representados por círculos para ilustrar uma unidade fundamental de processamento.

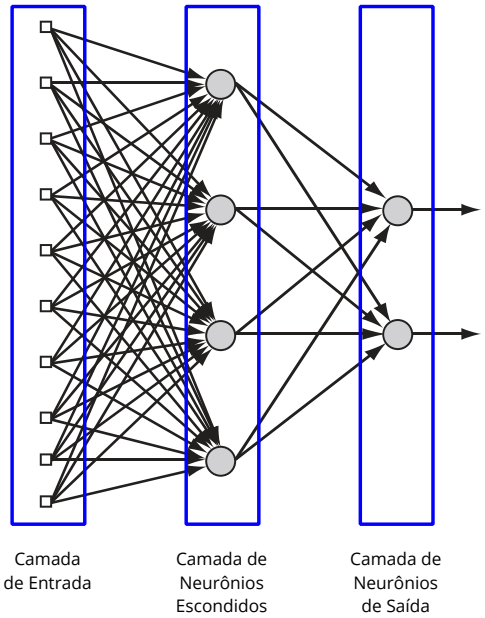
Da Figura 8 pode ser observado que a ligação entre neurônios em uma rede do tipo *feedforward* (“propagação à frente”) é feita por setas unidirecionais para representar que a direção do fluxo de informação é sempre da camada atual para a camada subsequente, nunca o inverso. Outra consideração em relação às conexões: quando todos os neurônios de uma camada se conectam com todos os neurônios da camada subsequente a rede é denominada completamente conectada, caso contrário, a rede será denominada parcialmente conectada. Sendo assim, as RNAs podem ser classificadas quanto ao número de camadas de processamento, direção do fluxo de informação e nível de conectividade das camadas (HAYKIN, 2009).

Haykin (2009) afirma que a principal propriedade de uma RNA é a sua capacidade de aprendizado. O mesmo autor define aprendizado no contexto de redes neurais como sendo um processo pelo qual os parâmetros livres de uma rede neural

Figura 8 – Estruturação de redes neurais em camadas.



(a) Rede Percptron de Camada Única



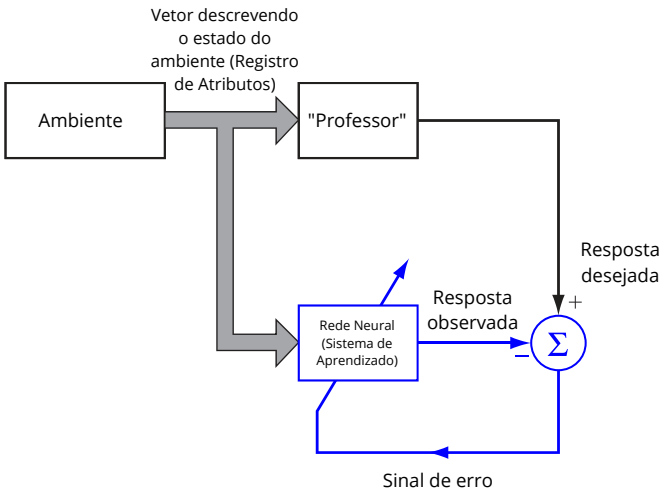
(b) Rede Perceptron Multicamadas

Fonte: Haykin (2009).

(pesos sinápticos e limiares) são adaptados através de um processo de estímulo pelo ambiente no qual a rede está inserida. Um tipo particular de processo de aprendizado é chamado de *Aprendizado Supervisionado*, onde dados previamente coletados que contêm valores para a variável que deseja-se aprender são utilizados como exemplos

de entrada e saídas desejadas para a rede durante o seu *treinamento*, Figura 9.

Figura 9 – Diagrama de blocos de um sistema de aprendizado supervisionado.



Fonte: Haykin (2009).

Cada registro contém os valores de todas as variáveis, inclusive da variável alvo. O processo ilustrado na Figura 9 ocorre da seguinte forma: os valores das variáveis que descrevem o problema (menos o da variável alvo) são apresentados à rede neural, que realiza uma estimativa da variável alvo. Após cada estimativa realizada é informada a *resposta correta* da variável alvo para que a rede tenha *"conhecimento"* do erro cometido e então, através de um algoritmo de aprendizado, os pesos sinápticos são corrigidos de maneira a diminuir a diferença entre o valor da variável alvo estimada e o seu valor real para as próximas estimativas (HAYKIN, 2009). Neste trabalho, tendo em vista a utilização de uma rede MLP, o algoritmo de treinamento utilizado foi o de retropropagação de erro ou *backpropagation*.

Tsung-Hsien, Chi-Kang et al. (2009) argumentam que não existe uma regra bem definida para a determinação do tamanho da amostra utilizada para treinamento de uma rede neural como modelo de previsão. Contudo, a prática têm apontado que uma amostra satisfatória de treinamento contêm registros suficientes para que a rede neural possa ter conhecimento das variáveis alvo para o período de previsão.

Tsung-Hsien, Chi-Kang et al. (2009) apresentam que, além da etapa de treinamento de uma rede neural, é necessário realizar a *validação* do processo de aprendizagem. A fase de validação tem por objetivo a determinação do nível de acurácia e precisão atingido pela rede nas previsões para dados não vistos durante o período de treinamento e consiste basicamente em analisar o grau de diferença entre a resposta observada e a resposta desejada (Figura 9). Sendo assim, a amostra total é dividida em duas partes: uma para treinamento (proporção maior) e outra para validação

(proporção menor) (TSUNG-HSIEN; CHI-KANG et al., 2009). A etapa de validação pode ser realizada, por exemplo, com base no cálculo do Erro Quadrático Médio (EQM), que é obtido através do somatório dos erros de previsão ao quadrado, divido pelo número de previsões realizadas, Equação (2.11a), e do Erro Médio Percentual Absoluto (EMPA), que é a média da diferença absoluta entre a resposta observada e a resposta desejada, expressa em porcentagem das respostas desejadas, Equação (2.11b).

$$EQM = \frac{1}{n}\sum_{i=1}^{i=n}(Ro_i - Rd_i)^2 \tag{2.11a}$$

$$EMPA = \frac{1}{n}\sum_{i=1}^{i=n}\left|\frac{Ro_i - Rd_i}{Rd_i}\right| \tag{2.11b}$$

Onde:

EQM - Erro Quadrático Médio;
$EMPA$ - Erro Médio Percentual Absoluto;
Ro_i - resposta observada i;
Rd_i - resposta desejada i.

Em relação à arquitetura de RNAs do tipo *Multilayer Perceptron* (MLP), isto é, número de camadas, bem como o número de neurônios em cada uma delas, Haykin (2009, p. 29) afirma ser um processo baseado em tentativa e erro, “infelizmente, atualmente não existem regras bem definidas para o processo de estruturação da rede”. Tsung-Hsien, Chi-Kang et al. (2009) sugerem que uma forma de determinar a arquitetura da rede é adicionar no mínimo um nó na camada de entrada para cada atributo explicativo da base de dados; não adicionar muitas camadas ocultas, e que uma forma de determinar o número de neurônios de uma camada oculta seria utilizar a média entre o número de neurônios da camada de saída e o número de nós da camada de entrada. Por exemplo, uma rede do tipo MLP construída para estimar a demanda de passageiros baseado em 5 atributos seria do tipo $5 - 3 - 1$; 5 nós de entrada, 3 neurônios na camada oculta e 1 neurônio na camada de saída.

As RNAs utilizadas neste trabalho são todas do tipo *Multilayer Perceptron* (MLP), completamente conetada, com uma camada oculta, que utilizam a Função Sigmóidal como função de ativação dos neurônios e o algoritmo *backpropagation* para ajustes dos pesos sinápticos.

2.4.2 Redes Neurais Artificiais como Modelos Previsão de Passageiros de Transporte

Como foi discutido anteriormente, RNAs possuem grande aptidão para funcionar como modelos de previsão. De especial interesse para este trabalho são os casos em que as RNA foram aplicadas como modelos de previsão de curto prazo da

demanda de passageiros em transporte. Como por exemplo em Dragana, Milica et al. (2015), Tsung-Hsien, Chi-Kang et al. (2009), Foell, Phithakkitnukoon et al. (2015), Zhang, Feng et al. (2016).

Dragana, Milica et al. (2015) reportou que a aplicação de uma Rede Perceptron Multicamadas $7 - 7 - 1$ completamente conectada alcançou uma média de 96% de precisão na previsão de vendas de bilhetes para três linhas ferroviárias da Sérvia, país do Leste Europeu, durante o período de férias escolares, permitindo a melhor adequação da oferta de lugares. No estudo realizado pelos autores, dados do ano de 2009 a 2013 sumarizados para os meses Junho, Julho, Agosto e Setembro foram utilizados. Cada registro de dado continha os seguintes atributos: a linha, o mês, o tipo de vagão, o número de lugares, o número de partidas, o preço do bilhete e o PIB do país para o mês em questão. A entrada era composta de 7 variáveis (atributos). Os autores optaram por 7 neurônios na camada escondida e 1 na camada de saída, já que a previsão era de apenas uma variável alvo.

Tendo em vista que o objetivo específico deste trabalho é apresentar uma RNA para previsão de demanda diária de passageiros de uma linha do transporte público por ônibus da Cidade de Joinville, o próximo capítulo tratará da caracterização do problema em questão bem como das características da RNA utilizada (função de ativação e arquitetura).

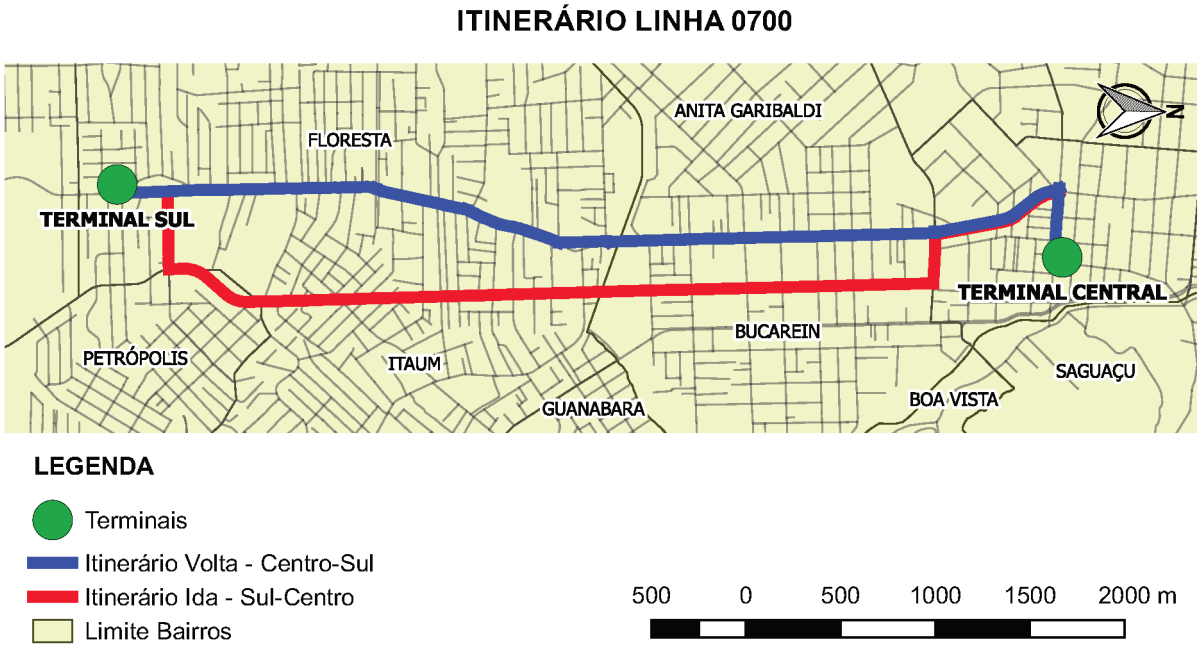
3 ESTUDO DE CASO

3.1 Caracterização do problema

Este trabalho tem por objetivo geral avaliar a aplicação de Redes Neurais Artificiais como modelos de produção (previsão) de viagens para auxílio no planejamento operacional do transporte público por ônibus. Por planejamento operacional entende-se a determinação da quantidade de lugares ofertados ao longo do dia, com o intuito de equilibrar a demanda por lugares no transporte e a oferta dos mesmos. Para tanto, uma linha de ônibus específica do transporte coletivo da Cidade de Joinville foi selecionada como estudo de caso para a aplicação de uma Rede percptron Multicamadas.

A Linha em questão é denominada 0700 Sul-Centro e foi escolhida por ser uma linha troncal relevante para o Sistema Integrado de Transporte de Joinville que conecta o Terminal da Região Sul da Cidade com o Terminal Central, Figura 10. A variável alvo (variável predita) é a *Quantidade de Passageiros por Dia*, e as variáveis explicativas são os demais atributos provenientes das bases descritas na Seção 3.2.

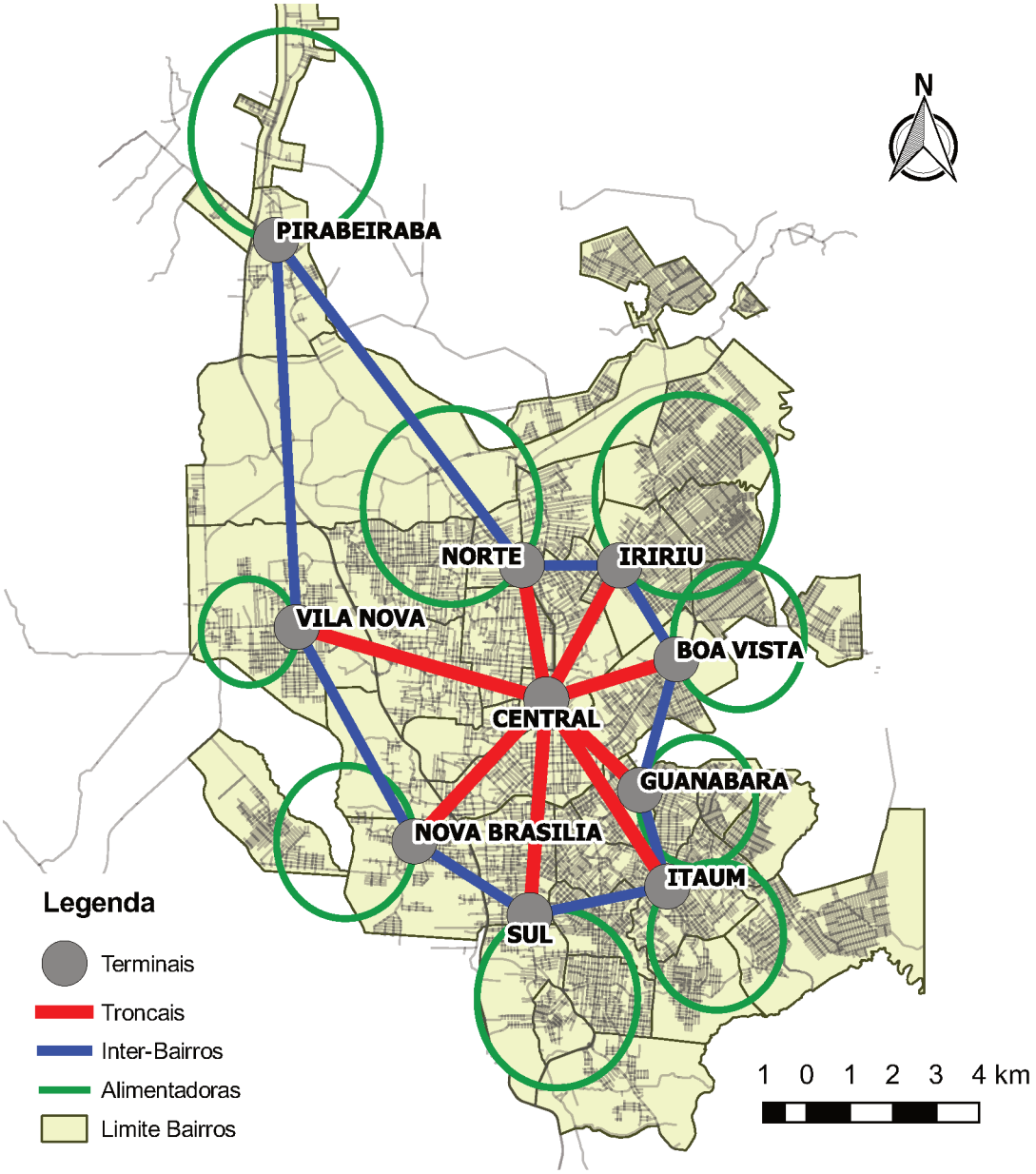
Figura 10 – Itinerário Linha 0700 Sul-Centro.



Fonte: Autor, 2018.

Atualmente, o deslocamento por ônibus é o principal modo de transporte coletivo público da Cidade de Joinville que é situada no Norte do Estado de Santa Catarina. Joinville é a cidade mais populosa do Estado. Segundo o IBGE (2018a), o número de habitantes em 2017 era de 577.077 pessoas. IPPUJ (2011) indica que, de todas as viagens realizadas dentro da Cidade, $26,48\%$ são através do transporte coletivo por ônibus. O Sistema de transporte público de Joinville é organizado de forma radial segundo classificação apresentada por Ferraz e Torres (2004), conforme Figura 11.

Figura 11 – Rede radial de transporte público na Cidade de Joinville.



Fonte: Autor, 2018.

De acordo com a Figura 11, o Sistema possui um terminal central que se conecta a outros oito terminais por meio de linhas troncais. A conexão entre terminais

não centrais é realizada por linhas inter-bairros perimetrais e de todos os terminais partem linhas alimentadoras que são responsáveis por distribuir as viagens dentro de cada região da Cidade.

Os terminais funcionam como estações de transbordo, uma vez que o Sistema de Transporte possui integração tarifária. Ou seja, se durante uma viagem, houver troca de linhas dentro dos terminais, o usuário não necessitará pagar outra tarifa. A única diferenciação tarifária presente no Sistema (além de descontos para estudantes municipais e isenções) é em relação ao momento de compra do bilhete. O valor da tarifa será maior, caso seja adquirido no ato de embarque em um ônibus; isto porque o Sistema não possui cobradores e o manuseio de tarifas fica sob responsabilidade do motorista, o que resulta no aumento do tempo total de embarque e, consequentemente, no aumento do tempo de viagem.

Dados provenientes da bilhetagem eletrônica do transporte coletivo, Figura 14, foram fornecidos pela Empresa Administradora do Sistema de Bilhetagem, denominada Passebus, para o período de $01/05/2015$ a $30/04/2017$. De acordo com a Seção 2.4.1 parte destes dados foram utilizados para o treinamento supervisionado da Rede Percptron Multicadas e os dados remanescentes para a validação do desempenho da Rede.

A Seção 3.2 apresenta de forma mais aprofundada o processo de aquisição dos dados, não só do sistema de bilhetagem eletrônica, mas de outras bases de dados da *Web*. A Seção 3.3 explica a forma como as bases de dados foram compatibilizadas e as anomalias removidas. A Seção 3.4 apresenta uma análise exploratória da variável alvo. A Seção 3.5 apresenta a seleção das variáveis explicatórias para o estudo, bem como a arquitetura desenhada para as redes utilizadas. Por último, a Seção 3.6 apresenta a forma como se deu o processo de aprendizagem e validação das redes.

3.2 Aquisição dos dados

Segundo Tan, Steinbach e Kumar (2005) a mineração de dados é parte substancial do processo de extração de conhecimento em bases de dados, do inglês, *Knowledge Discovery in Databases* (KDD), Figura 12, que é o processo encarregado da conversão de dados *brutos* em informação útil.

De acordo com Figura 12, o processo de KDD consiste em uma série de etapas de transformações que vai do pré-processamento ao pós-processamento, passando pela mineração de dados. Contudo, existe uma etapa prévia à etapa de pré-processamento, representada pelo bloco “Aquisição de Dados” na Figura 12, cuja função é reunir dados relevantes que serão utilizados para gerar conhecimento em relação a um problema. Dados podem ser armazenados de diversas formas em diferentes localidades como, por exemplo, em arquivos de textos, tabelas relacionais e

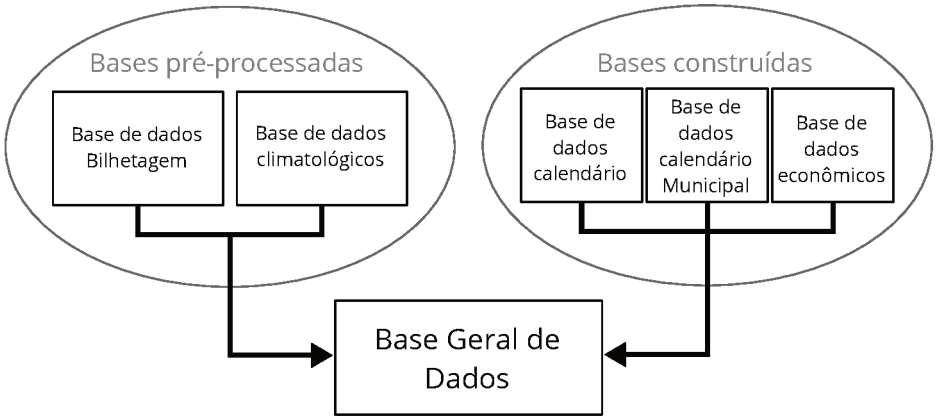
Figura 12 – Processo de extração de conhecimento em bases de dados (KDD).

Fonte: Tan, Steinbach e Kumar (2005).

planilhas.

Esta seção visa descrever o processo de aquisição, e as características, dos dados utilizados para construção da "Base Geral de Dados", que é composta pela quantidade de passageiros por dia (variável alvo) da Linha 0700 e outras 19 variáveis independentes oriundas de 5 diferentes bases de dados, conforme esquema na Figura 13. A Base Geral de Dados serviu para treinamento e validação da Rede Perceptron Multicamadas estruturada para o problema da previsão do número de passageiros por dia na Linha 0700 do Transporte Público de Joinville.

Figura 13 – Construção da Base Geral de Dados



Fonte: Autor, 2018.

A base denominada Base Geral de Dados é resultado da junção de cinco diferentes bases de dados, conforme Figura 13, sendo três destas construídas pelo Autor, uma fornecida pela Empresa Administradora do Sistema de Bilhetagem e uma disponível no *site* da Prefeitura Municipal de Joinville. A Base Geral de Dados possui os seguintes atributos: Ano_Saida, Mes_Saida, Dia_Saida, Dia_Semana_Saida, Semana_Mes_Saida, Semana_Do_Ano, Vespera_Feriado, Feriado, Feriado_Descricao, Pos_Feriado, Recesso_Escolar, Festivais, Festivais_Descricao, Preco_Gasolina,

Preco_Passagem_Antecipada, Preco_Passagem_Embarcada, Temperatura_Maxima, Temperatura_Minima, Chuva, Sentido e Passageiros (variável alvo). A variável "Passageiros" representa o total de usuários ao longo do dia para a Linha 0700.

O processo de aquisição e as características dos dados de cada uma das bases de dados utilizadas para a construção da "Base Geral de Dados" é descrito nas sub-seções que seguem.

3.2.1 Dados do transporte coletivo

A principal base de dados para o presente estudo foi fornecida pela Empresa Administradora do Sistema de Bilhetagem Eletrônica, denominada Passebus. O Transporte Coletivo de Joinville possui bilhetagem eletrônica através de *smart cards*, Figura 14a. O sistema de bilhetagem é composto por um conjunto leitor-catraca, Figura 14b, e por uma unidade de processamento, Figura 14c, que tem por função registrar as informações das viagens e é operada pelo condutor do veículo.

Figura 14 – Sistema de bilhetagem eletrônica.

(a) *Smart card* (b) Validador (c) Unidade de processamento

Fonte: Autor, 2018.

Ao final de cada viagem a unidade de processamento faz o registro dos seguintes atributos: Código da Empresa, Identificador do Veículo, Identificador da Linha, Data de Saída da Viagem, Hora de Saída, Data de Chegada da Viagem, Hora de Chegada, Sentido da Viagem, Trecho, Quantidade de Passageiros na Catraca e a Quantidade de Passageiros Embarcados. O tipo de cada uma das variáveis citadas anteriormente, bem como a forma de registro é descrita a seguir:

- **Código da Empresa**: Variável que identifica a operadora responsável pela viagem em questão. Entrada registrada automaticamente pela unidade de processamento. Considerando-se que existem duas operadoras de transporte em Joinville, essa variável pode assumir os valores 01 ou 02;
- **Identificador do Veículo**: Variável que identifica o veículo utilizado para a realização da viagem. Entrada registrada automaticamente pela unidade de processamento. Variável composta somente por números, como exemplo: 15202;
- **Identificador da Linha**: Variável que identifica a linha da viagem. Entrada

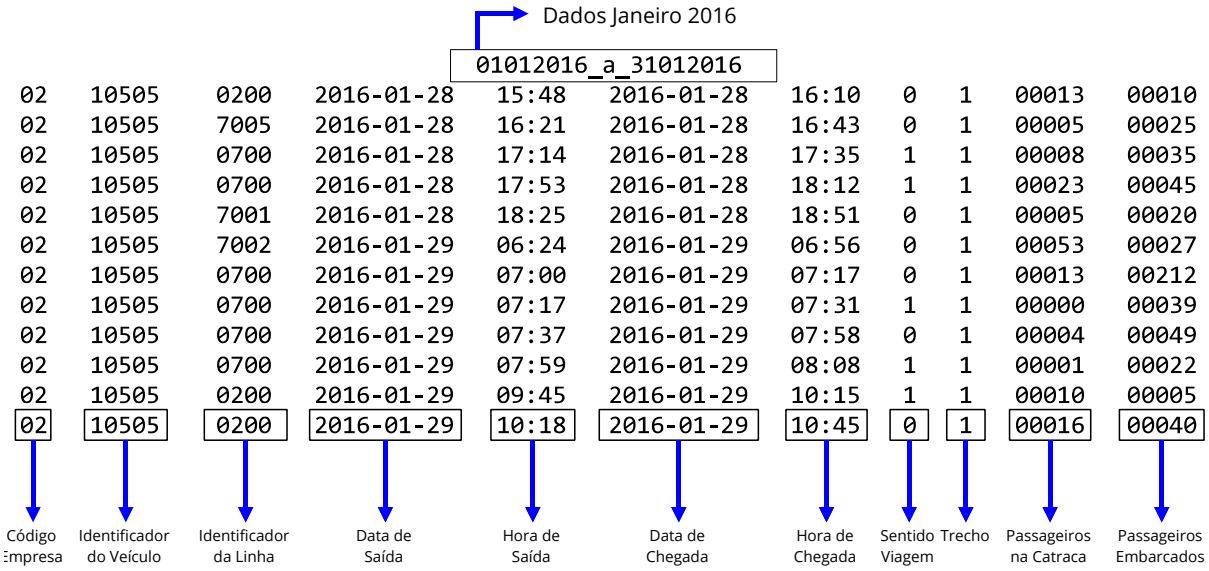
registrada pelo motorista. Variável composta somente por números, como exemplo: 0700;

- **Data de Saída**: Variável que indica a data de partida da viagem. Entrada registrada pelo motorista. Variável composta pelo dia do mês, mês e ano separados por "/", exemplo $21/05/2015$;
- **Hora de Saída**: Variável que indica a hora de partida da viagem. Entrada registrada pelo motorista. Variável composta pela hora, variando de 00 a 23, e minutos, variando de 00 a 59, separados por ":", exemplo $17:59$;
- **Data de Chegada**: Variável que indica a data de chegada da viagem. Entrada registrada pelo motorista. Variável composta pelo dia do mês, mês e ano separados por "/";
- **Hora de Chegada**: Variável que indica a hora de chegada da viagem. Entrada registrada pelo motorista. Variável composta pela hora, variando de 00 a 23, e minutos, variando de 00 a 59, separados por ":";
- **Sentido da Viagem**: Variável que identifica o sentido da viagem em questão. Entrada registrada pelo motorista. Esta variável pode assumir os valores 0 para indicar "ida" ou 1 para indicar "volta";
- **Trecho**: atributo registrado automaticamente sempre com o valor 1;
- **Passageiros na Catraca**: Variável que indica a quantidade de passageiros que passaram pela catraca entre a Hora de Saída e a Hora de Chegada. Entrada registrada automaticamente pela unidade de processamento. Variável discreta que, teoricamente, pode assumir qualquer valor dentro do conjunto $\mathbb{N}$;
- **Passageiros Embarcados**: Variável que indica a quantidade de passageiros que entram pelas portas de desembarque do ônibus, quando o mesmo se encontra parado nos terminais. Isto é, esta variável representa o número de pessoas que estão presentes dentro do ônibus antes que a viagem tenha início. Entrada registrada pelo motorista através de contagem manual. Variável que, teoricamente, pode assumir qualquer valor dentro do conjunto $\mathbb{N}$.

Os dados da Bilhetagem são exportados do Sistema em formato de "tabelas" sem cabeçalhos, agrupadas mensalmente. Cada linha das tabelas presentes nos arquivos representa uma viagem, e cada coluna representa um dos atributos descritos anteriormente, Figura 15.

Como pode ser observado na Figura 15, a base de dados do transporte coletivo é a de maior relevância para o presente estudo por apresentar os valores da variável alvo. A variável "Passageiros" na "Base Geral de Dados" é resultado do somatório das variáveis "Passageiros na Catraca" e "Passageiros Embarcados" como será explicado na Seção 3.3.

Figura 15 – Representação arquivo “.File” proveniente do Sistema de Bilhetagem.



Fonte: Autor, 2018.

3.2.2 Dados de calendário

Baseado na característica de análise multi-hipóteses dos métodos de mineração de dados, Seção 2.4, a adição de atributos, além daqueles que estão presentes na base de dados do transporte público, tem o intuito de viabilizar a análise em relação a outras variáveis explicativas que podem ter relação com a variação do número de usuários ao longo da semana, do mês, do ano, e assim por diante.

Desta forma, para o intervalo de tempo dos dados do transporte público, $01/05/2015$ a $30/04/2017$, foi construída uma base de dados, em formato “.CSV”, contendo os seguintes atributos: Ano, Mes, Dia, Dia_Semana, Semana_Do_Mes, Semana_Do_Ano, Vespera_Feriado, Feriado, Feriado_Descricao e Pos_Feriado, Tabela 2.

Tabela 2 – Base de dados Calendário

Ano	Mes	Dia	Dia_Semana	Semana_Do_Mes	Semana_Do_Ano	Vespera_Feriado	Feriado	Feriado_Descricao	Pos_Feriado
2015	5	1	6	1	17	0	1	Dia_Do_Trabalho	0
2015	5	2	7	1	17	0	0	NA	1
2015	5	3	1	1	18	0	0	NA	0
2015	5	4	2	1	18	0	0	NA	0
2015	5	5	3	1	18	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2015	5	14	5	2	19	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2017	4	29	7	5	17	0	0	NA	0
2017	4	30	1	5	18	1	0	NA	0

Fonte: Autor, 2018.

As informações de Ano, Mês, Dia, Dia_Da_Semana, Semana_Do_Mes e Semana_Do_Ano foram extraídas do Softaware Excel®. Já os dados de

Vespera_Feriado, Feriado, Feriado_Descricao e Pos_Feriado podem ser encontradas em Brasil (2015), Brasil (2016) e Brasil (2017). O tipo de cada uma das variáveis presentes na base de dados de calendário é descrito a seguir:

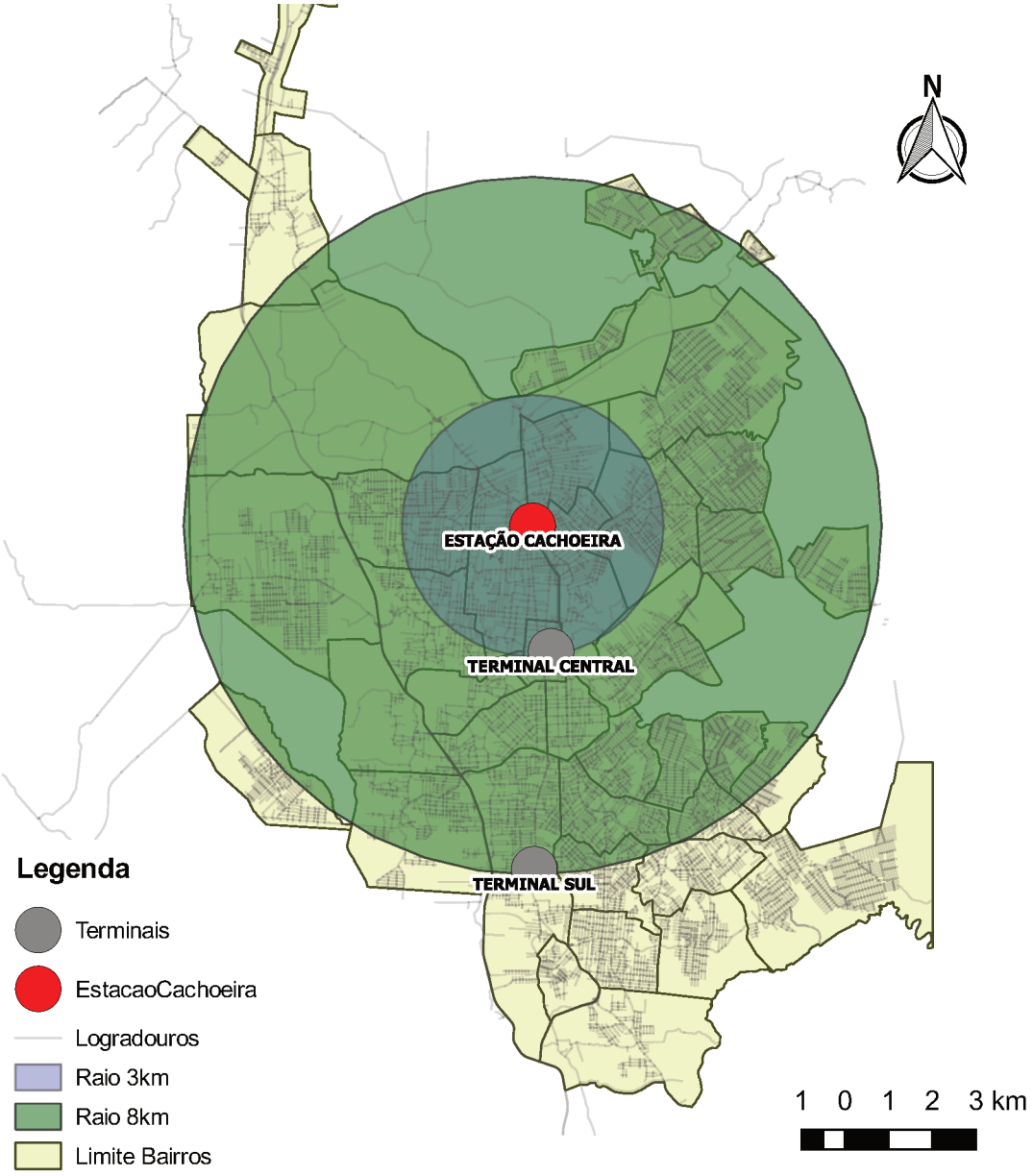
- **Ano**: Variável que indica o ano do registro. A variável "Ano" pode assumir um dos seguintes valores: 2015, 2016 ou 2017;
- **Mês**: Variável que indica o mês do registro. A variável "Mês" pode assumir valores entre 1 e 12;
- **Dia**: Variável que indica o dia do registro. A variável "Dia" pode assumir valores entre 1 e 31;
- **Dia_Da_Semana**: Variável que identifica o dia da semana do registro, como por exemplo, Segunda-feira, Terça-feira e assim sucessivamente. A variável "Dia_Da_Semana" pode assumir valores entre 1 e 7, sendo que o valor 1 representa o Domingo, 2 a Segunda-feira, 3 a Terça-feira, 4 a Quarta-feira, 5 a Quinta-feira, 6 a Sexta-feira, e o valor 7 representa o Sábado;
- **Vespera_Feriado**: Variável que indica se o dia do registro é um dia que antecede algum feriado. A variável "Vespera_Feriado" pode assumir os valores 0 para indicar "falso" ou 1 para indicar "verdadeiro";
- **Feriado**: Variável que indica se o dia do registro é um feriado. A variável "Feriado" pode assumir os valores 0 para indicar "falso" ou 1 para indicar "verdadeiro";
- **Feriado_Descricao**: Variável que identifica o feriado em questão. A variável "Feriado_Descricao" pode assumir valores textuais (*strings*) como por exemplo, "Primeiro_Do_Ano";
- **Pos_Feriado**: Variável que indica se o dia do registro é um dia que sucede algum feriado. A variável "Pos_Feriado" pode assumir os valores 0 para indicar "falso" ou 1 para indicar "verdadeiro".

3.2.3 Dados climatológicos

A base de dados climatológicos foi extraída da Rede de Monitoramento das Estações Meteorológicas de Joinville, disponível ao cidadão mediante de cadastro em Joinville (2017b), e contém o registro diário de dados climáticos para o período de $19/04/2012$ a $28/07/2017$, medidos a partir da Estação Hidrometeorológica denominada Cachoeira Área Central, Figura 16.

As seguintes variáveis compõem a base de dados climatológicos: TIMESTAMP, RECORD, Temp_Ar_Max, Temp_Ar_TMx, Temp_Ar_Min, Temp_Ar_TMn, Umid_Rel_Max, Umid_Rel_TMx, Umid_Rel_Min, Umid_Rel_TMn, Rad_Total_Tot, Rajada, Dir_Rajada, Chuva_Tot, Nivel_Max, Nivel_TMx, Nivel_Min, Nivel_TMn, Orvalho, Ind_Calor e WindChill. Todas as variáveis terminadas em "TMx" ou "TMn" representam a hora de ocorrência (T = *timestamp*) da variável máxima e mínima respectivamente.

Figura 16 – Localização Estação Hidrometeorológica Cachoeira Área Central.



Fonte: Autor, 2018.

Por exemplo, a variável “Temp_Ar_TMx” apresenta a hora de ocorrência da máxima temperatura atingida em determinado dia. Para efeitos de entendimento será realizado apenas a descrição das variáveis que foram adicionadas na “Base Geral de Dados”:

- **TIMESTAMP**: Variável que indica a data, hora e período do dia do registro. A variável “TIMESTAMP” é do formato *data-hora-período*, como por exemplo, $28-07-201712:00:00AM$.
- **Temp_Ar_Max**: Variável que indica a máxima temperatura atingida em determinado dia. A variável “Temp_Ar_Max” é dada em $^{\circ}C$ pode assumir qualquer valor dentro da escala Celsius;

- **Temp_Ar_Min**: Variável que indica a máxima temperatura atingida em determinado dia. A variável “Temp_Ar_Max” é dada em $^{\circ}C$ pode assumir qualquer valor dentro da escala Celsius;
- **Chuva_Tot**: Variável que indica a quantidade total de chuva, em milímetros, em determinado dia. A variável “Chuva_Tot” pode assumir qualquer valor dentro dos $\mathbb{R}$.

A Seção 3.3 explicará como foi realizada a extração dos atributos utilizados da base de dados climatológicos e a compatibilização destes atributos com a “Base Geral de Dados”.

3.2.4 Dados do calendário Municipal: recesso escolar e festival de dança

Segundo Ferraz e Torres (2004), a educação é um dos maiores motivos pelos quais as pessoas se locomovem. Desta forma, acrescentar um campo de atributo na “Base Geral de Dados” para identificar se determinado dia corresponde a um dia de recesso letivo permite avaliar a influência do período de férias na demanda por transporte público. Apenas o período de recesso Municipal é considerado partindo da premissa que o calendário de Instituições de ensino estaduais, federais e particulares possuem calendário letivo semelhante.

A Cidade de Joinville realiza, anualmente, o maior festival de dança do planeta em número de participantes. Recorde detido pela Cidade desde o ano de 2005 no *Guinness Book* (JOINVILLE, 2017d). O Festival geralmente ocorre na segunda metade do mês de Julho e tem duração de aproximadamente 12 dias. Baseado na hipótese de que existe influência na demanda de passageiros do transporte público, devido ao grande número de visitantes do Festival, um campo de atributo foi acrescentado na “Base Geral de Dados” para identificar se determinado dia corresponde a um dia de Festival.

A base de dados do calendário Municipal foi construída em formato “.CSV” a partir de Joinville (2015a), Joinville (2016), Joinville (2017a) para os dados referentes ao calendário letivo das Escolas Municipais e Joinville (2017d) para os dados relativos ao Festival, para o período de $01/05/2015$ a $30/04/2017$, e contém os seguintes atributos: Ano, Mês, Dia, Recesso_Escolar, Festivais e Festivais_Descricao, Tabela 3. As variáveis “Ano”, “Mês” e “Dia” já foram descritas anteriormente (vide seção 3.2.2). Sendo assim, somente os tipos das demais variáveis serão descritos a seguir:

- **Recesso_Escolar**: Variável que indica se o dia do registro é um dia de recesso escolar Municipal. A variável “Recesso_Escolar” pode assumir os valores 0 para indicar “falso” ou 1 para indicar “verdadeiro”;

Tabela 3 – Base de dados Calendário Municipal

Ano	Mes	Dia	Recesso_Escolar	Festivais	Festivais_Descricao
2015	5	1	0	0	NA
⋮	⋮	⋮	⋮	⋮	⋮
2016	7	27	1	1	Festival_De_Danca
2016	7	28	1	1	Festival_De_Danca
2016	7	29	1	1	Festival_De_Danca
2016	7	30	1	1	Festival_De_Danca
2016	7	31	1	0	NA
2016	8	1	0	0	NA
⋮	⋮	⋮	⋮	⋮	⋮
2017	4	30	0	0	NA

Fonte: Autor, 2018.

- **Festivais**: Variável que indica se o dia do registro é um dia de festival. A variável "Festivais" pode assumir os valores 0 para indicar "falso" ou 1 para indicar "verdadeiro";
- **Festivais_Descricao**: Variável que identifica o festival em questão. A variável "Festival_Descricao" pode assumir valores textuais (*strings*) como por exemplo, "Festival_De_Danca".

3.2.5 Dados econômicos: preço da passagem de ônibus e litro da gasolina

Levando-se em consideração que o custo financeiro associado ao deslocamento por determinado modal influencia diretamente na sua Função Utilidade, Seção 2.2.3, e consequentemente na probabilidade de escolha do modal em questão. Decidiu-se por acrescentar, à "Base Geral de Dados", o preço de 1 passagem de ônibus para determinado dia, que corresponde ao custo monetério de uma viagem, e o preço de 1 litro de gasolina, que é o combustível utilizado pela maioria dos automóveis de passeio, principal concorrente do ônibus. O objetivo da adição desses dois atributos na "Base Geral de Dados" é permitir testar a hipótese de que variações entre o preço da passagem e da gasolina influenciam na demanda de passageiros do ônibus.

A base de dados econômicos foi construída, em formato ".CSV',' para o período de $01/05/2015$ a $30/04/2017$, a partir das informações presentes em ANP (2018), para o levantamento do preço médio semanal do litro da gasolina para o Município de Joinville, e Joinville (2014), Joinville (2015b) e Joinville (2017c), para as informações referentes à mudança do preço da passagem de ônibus.

ANP (2018) apresenta o preço médio semanal do litro da gasolina. Deste modo, para todos os dias dentro de uma semana, na base de dados econômicos, o

preço da gasolina foi admitido constante. Considerando-se que o preço da passagem sofre alteração dependendo do momento de aquisição da mesma, a base de dados econômicos contém os seguintes atributos: Ano, Mês, Dia, Preco_Gasolina, Preco_Passagem_Antecipada e Preco_Passagem_Embarcada, Tabela 4. As variáveis "Ano", "Mês" e "Dia" já foram descritas anteriormente (vide seção 3.2.2). Sendo assim, os tipos das demais variáveis serão descritos a seguir:

- **Preco_Gasolina**: Variável que indica o preço médio semanal de 1 litro de gasolina em $R\$$. A variável "Preco_Gasolina" pode assumir qualquer valor dentro dos $\mathbb{R}$;
- **Preco_Passagem_Antecipada**: Variável que indica o preço de 1 bilhete adquirido antes do momento de embarque no ônibus em determinado dia. A variável "Preco_Passagem_Antecipada" pode assumir qualquer valor dentro dos $\mathbb{R}$;
- **Preco_Passagem_Embarcada**: Variável que indica o preço de 1 bilhete adquirido no momento de embarque no ônibus em determinado dia. A variável "Preco_Passagem_Embarcada" pode assumir qualquer valor dentro dos $\mathbb{R}$;

Tabela 4 – Base de dados Informações Econômicas

Ano	Mes	Dia	Preco_Gasolina	Preco_Passagem_Antecipada	Preco_Passagem_Emba
2015	5	1	3,013	3,25	3,7
⋮	⋮	⋮	⋮	⋮	⋮
2016	7	27	3,267	3,7	4,5
2016	7	28	3,267	3,7	4,5
2016	7	29	3,267	3,7	4,5
2016	7	30	3,267	3,7	4,5
2016	7	31	3,286	3,7	4,5
⋮	⋮	⋮	⋮	⋮	⋮
2017	4	30	3,272	4,0	4,5

Fonte: Autor, 2018.

3.3 Pré-processamento dos dados

Segundo Tan, Steinbach e Kumar (2005) a etapa de pré-processamento diz respeito à junção de dados provenientes de múltiplas fontes, seleção dos registros e atributos relevantes para o problema em questão e remoção de anomalias. Os mesmos autores argumentam que a etapa de pré-processamento é o passo mais trabalhoso e demorado dentro do processo de extração de conhecimento em bases de dados (KDD), Figura 12, devido ao fato de que existem diversos procedimentos para coleta e armazenamento de dados.

3.3.1 Seleção de registros das bases de dados e remoção de anomalias

Dentro do escopo deste trabalho, a etapa de pré-processamento teve início com a remoção de registros da base de dados do transporte público. Primeiramente, foram removidos os registros que não correspondiam as viagens realizadas pela Linha 0700. Depois, foram removidos os registros que continham erros óbvios de digitação para a Variável “Passageiros Embarcados”.

Visto que a Linha 0700 tem início sempre no Terminal Sul ou no Terminal Central (dependendo do sentido), e que nos terminais os passageiros embarcam pelas portas traseiras sem passarem por qualquer tipo de controle, como uma catraca por exemplo, a contagem do número de passageiros dentro do veículo antes do início da viagem é realizada manualmente pelo motorista e então registrada na unidade de processamento da bilhetagem eletrônica. Por ser um procedimento manual, a contagem e o registro da variável “Passageiros Embarcados” podem conter erros. Pouco se pode afirmar sobre a acurácia da contagem realizada pelo motorista, contudo, alguns erros de digitação são facilmente identificáveis, Figura 17(a). Isto porque os ônibus possuem uma capacidade estipulada e as Concessionárias são legalmente proibidas de realizar viagens com o número de passageiros acima dessa capacidade, sob pena de multa.

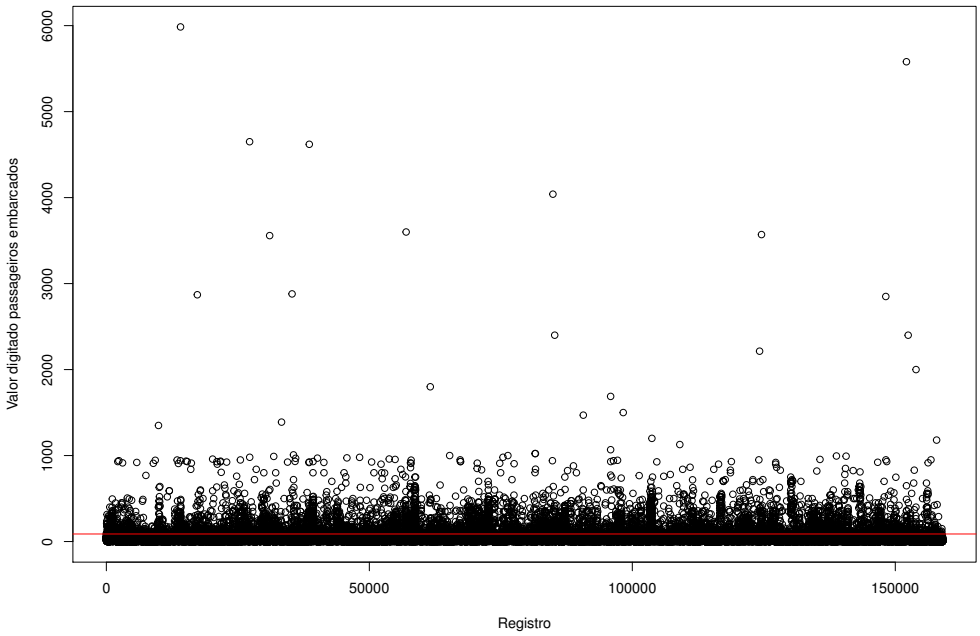
Sendo assim, utilizando como limitante o valor da capacidade dos veículos, todos os registros da base de dados do transporte público que possuíam um valor maior para a variável “Passageiros Embarcados” do que a capacidade do ônibus presente no registro, tiveram o valor da variável “Passageiros Embarcados” substituído pela média arredondada para cima do valor de “Passageiros Embarcados” dos 159.006 registros presentes na base do transporte público para a linha 0700, sendo o valor da média de 41 passageiros embarcados, Figura 17(b). Existiam 7.997 registros com erro de digitação, o equivalente a 5% do total de registros de viagens.

Como pode ser observado na Figura 17(a), existiam registros da ordem de milhares para a variável “Passageiros Embarcados”. Após o tratamento realizado, o maior valor encontrado é de 93 passageiros que é exatamente igual à capacidade do maior ônibus utilizado para a realização da Linha 0700, 17(b). A Reta $Valor\ digitado\ passageiros\ embarcados = 93$ é representada pela linha vermelha nas Figuras 17(a) e 17(b).

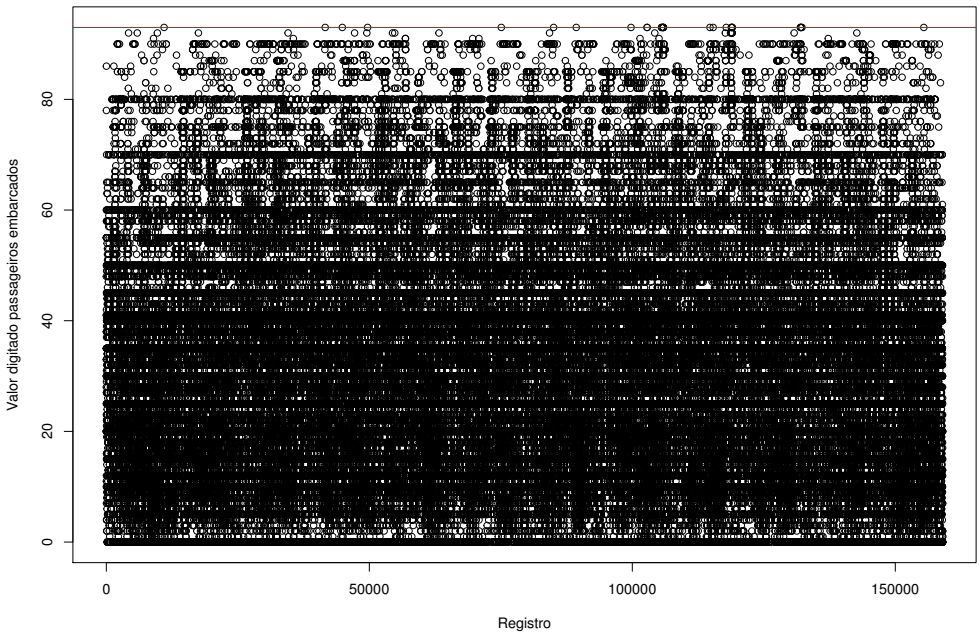
O total de passageiros para uma viagem qualquer é o resultado da soma entre o número de “Passageiros Embarcados” e o número de “Passageiros Catracados”, desta forma, para cada registro da base de dados da Linha 0700 foi criado um novo atributo denominado “Passageiros” que representa o total de passageiros em uma viagem.

Posto que a variável a ser predita é a *Quantidade de Passageiros por Dia* houve a necessidade de agregar o número da variável “Passageiros” de todas as viagens da

Figura 17 – Seleção de registros.



(a) Anomalias no registro da variável Passageiros Embarcados.



(b)Variável Passageiros Embarcados após remoção de anomalias.

Fonte: Autor, 2018.

Linha 0700 para um mesmo dia e para um mesmo sentido em um único registro. A base de dados resultante desse tratamento de registros apresenta para cada dia dentro do intervalo de $01/05/2015$ a $30/04/2017$ o número total de passageiros ao longo do dia

para a linha 0700.

Após a etapa de seleção de registros da base de dados do transporte público, optou-se por realizar o mesmo procedimento na base de dados climatológicos. De acordo com a Seção 3.2.3 a base de dados climatógicos apresenta o registro diário de diversas variáveis de clima e se extende do período de $19/04/2012$ a $28/07/2017$. Para efeitos de compatibilização com a base do transporte público, em termos de datas, foram removidos os registros de $19/04/2012$ a $30/04/2015$ e depois de $01/05/2017$ a $28/07/2017$ restando apenas o intervalo de $01/05/2015$ a $30/04/2017$.

3.3.2 Seleção de atributos das bases de dados

Os atributos removidos da base dados do transporte público foram: “Código da Empresa” e “Trecho”, visto que as variáveis assumem valores constantes dentro da base de dados, 02 e 1, respectivamente; “Identificador do Veículo”, “Hora de Saída” e “Hora de Chegada”, dado que os passageiros foram agrupados por dia, e “Data de Chegada”, pelo fato de que o as RNAs construídas para este trabalho são Modelos de Produção de Viagens na origem. Deste modo, a base de transporte público resultante da etapa de seleção de atributos retringe-se aos seguintes atributos: “Data de Saída”, “Sentido” e “Passageiros” que apresenta a quantidade de passageiros ao longo de um determinado dia para a Linha 0700.

Em relação à base de dados climatógicos os únicos atributos considerados relevantes foram aqueles que dizem respeito à temperatura e chuva, partindo da premissa que de todas as variáveis apresentadas na base em questão, como por exemplo, umidade do ar, radiação solar e nível da maré, a temperatura e a chuva são as características climáticas que mais influenciam na escolha modal. Por isso, a base climatológica resultante da etapa de seleção de atributos limita-se às variáveis “TIMESTAMP” que apresenta o dia do registro, “Temp_Ar_Max”, temperatura máxima para o dia em questão, “Temp_Ar_Min”, temperatura mínima para o dia em questão, e “Chuva_Tot” que representa o volume pluviométrico do dia.

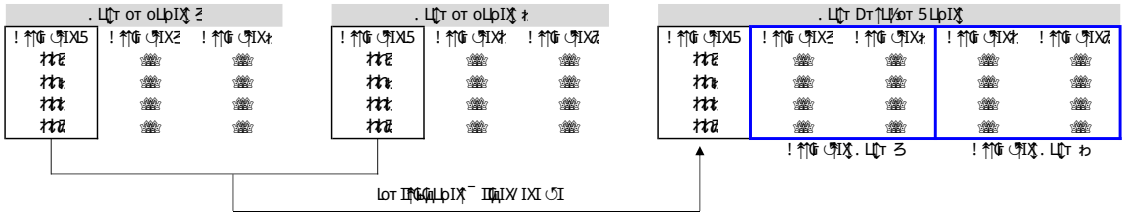
3.3.3 Junção das bases de dados

Como foi apresentado na Seção 3.2 a “Base Geral de Dados” é resultado da junção de 5 diferentes bases de dados. Para efetuar a união de bases de dados de múltiplas fontes foi adotado o método de *Identificador Único Comum* apresentado em Longley et al. (2013), que consiste em agregar em uma única base, atributos provenientes de outras bases de dados que são referentes ao mesmo identificador único.

A Figura 18, retrata o conceito de identificador único comum na qual a “Base de dados 1” possui um registro que contêm valores quaisquer de “Atributo 1” e “Atributo

2" para o elemento que apresenta "Atributo ID $= 001$" e a "Base de dados 2" possui um registro que contêm valores quaisquer de "Atributo 3" e "Atributo 4" para o elemento de "Atributo ID $= 001$", sendo assim, é possível na, "Base Geral de Dados", construir um registro contendo os Atributos 1, 2, 3 e 4 para o identificador único comum *Atributo ID* 001.

Figura 18 – Esquema de junção de bases de dados através de ID Único Comum.



Fonte: Autor, 2018.

Como a variável a ser predita é a *Quantidade de Passageiros por Dia* decidiu-se por utilizar como identificador único comum o conjunto *Dia, Mês, Ano* que está presente em todas as 5 bases.

Como foi apresentado na Seção 3.2, a Base Geral de Dados possui 731 registros (quantidade de dias presentes no intervalo de $01/05/2017$ a $30/04/2017$), e os seguintes atributos: Ano_Saida, Mes_Saida, Dia_Saida, que formam o conjunto de identificador comum único; Dia_Semana_Saida, Semana_Mes_Saida, Semana_Do_Ano, Vespera_Feriado, Feriado, Feriado_Descricao, Pos_Feriado, proveniente da base de calendário; Recesso_Escolar, Festivais, Festivais_Descricao, provenientes da base calendário municipal; Preco_Gasolina, Preco_Passagem_Antecipada, Preco_Passagem_Embarcada, provenientes da base de dados econômicos; Temperatura_Maxima, Temperatura_Minima, Chuva, proveninetes da base climatológica; Sentido e Passageiros (variável alvo) proveniente da base do transporte público. A variável "Passageiros" representa o total de usuários ao longo do dia para a Linha 0700.

3.4 Análise exploratória da Variável Alvo

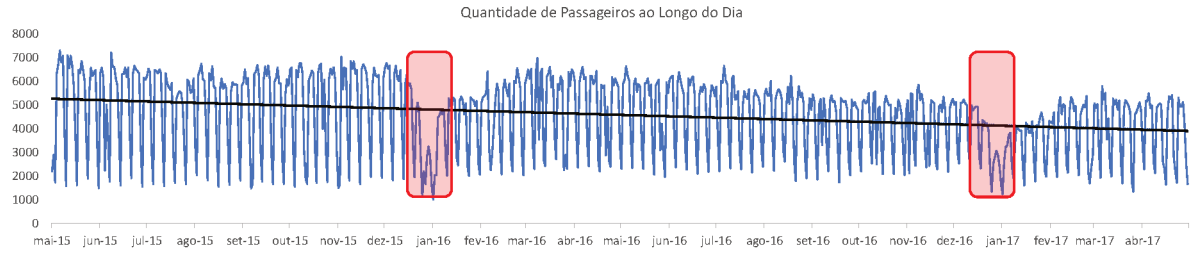
A Variável alvo "Quantidade de Passageiros ao Longo do Dia", presente na "Base Geral de Dados", pode ser classificada como uma série temporal uma vez que os registros presentes na base são ordenados sequencialmente no tempo (Figura 19). Makridakis, Wheelwright e Hyndman (1997) afirmam que uma série temporal possui quatro componentes principais:

- **Tendência**: indica a direção contínua na qual os dados estão se movimentando. É avaliada em períodos longos de tempo;

- **Ciclo**: movimento oscilatório dos dados em torno da tendência, que pode ser observado durante períodos de tempo maiores que um ano e que não são necessariamente regulares;
- **Sazonalidade**: movimentos ocilatórios que ocorrem com regularidade em períodos de tempo fixo, como por exemplo, semanas, meses e trimestres;
- **Componente Aleatório**: ocilações causadas nos dados por fenômenos aleatórios.

Na Figura 19 é possível observar que a série temporal que representa a variável "Quantidade de Passageiros ao Longo do Dia" para o período analisado neste trabalho ($01/05/2015$ a $30/04/2017$) possui uma componente de tendência (linha reta de cor preta) negativa, o que expressa que o número de passageiros na Linha 0700 vêm decaindo ao longo do tempo. Também na Figura 19 fica evidente que o período de férias escolares de verão (assinalado pelas áreas de borda vermelha) afetam significativamente a demanda por transporte.

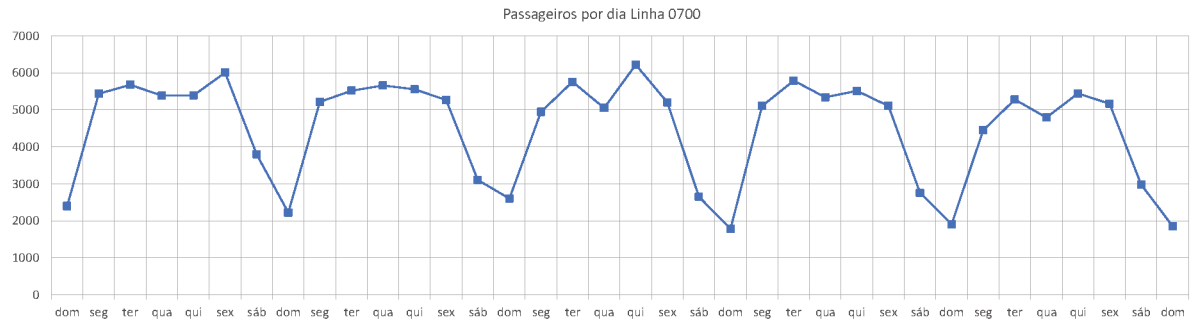
Figura 19 – Quantidade de passageiros por dia Linha 0700 como uma série temporal.



Fonte: Autor, 2018.

A Figura 20 retrata a sazonalidade semanal presente na utilização da Linha 0700 (objeto de estudo do presente trabalho). Quatro semanas foram escolhidas aleatóriamente e plotadas em sequência para que se pudesse observar a diferença presente no número de passageiro para dias úteis e finais de semana.

Figura 20 – Sazonalidade da quantidade de passageiros por dia Linha 0700.



Fonte: Autor, 2018.

A análise exploratória da existência de componentes de tendência e sazonais no transporte coletivo é de grande importância, pois a decisão de inclusão ou não de atributos explicativos para essas componentes influencia diretamente o processo de escolha de variáveis fornecidas à rede neural, como será tratado na Seção 3.5.

3.5 Arquitetura das Redes Neurais utilizadas

De acordo com a Seção 2.4.1, a arquitetura de uma RNA diz respeito às seguintes características: quantidade de camadas, número de neurônios em cada camada, grau de conexão entre os neurônios de camadas adjacentes e direção do fluxo de informação.

Segundo Tan, Steinbach e Kumar (2005), RNAs com apenas uma camada oculta são estimadores universais que funcionam bem para a maioria dos problemas de classificação e regressão. Por este motivo, todas as redes neurais utilizadas nesse trabalho possuem apenas uma camada oculta, ou seja, são do tipo $p - h_1 - q$, o que é equivalente a afirmar que as RNAs possuem uma camada de entrada com p nós de entradas, uma camada oculta com h_1 neurônios e uma camada de saída com q neurônios.

A determinação da quantidade de unidades em cada camada varia de acordo com a camada em questão. Para a camada de entrada, o número p de nós de entrada corresponde à quantidade de atributos explicativos presentes na base de dados utilizada para treinamento. Por sua vez, na camada de saída, o número q de neurônios é determinado pela quantidade de variáveis alvo que se pretende estimar, que no caso deste trabalho será sempre igual a 1. Por último, para a camada oculta a determinação do número h_1 "ótimo" de neurônios se dá de modo empírico.

Apesar de a "Base Geral de Dados" possuir potencialmente 19 atributos explicativos, optou-se por não utilizá-los todos juntos como parâmetros de entrada da rede. Foram definidos 6 subconjuntos de atributos para que se pudesse avaliar variações na acurácia da previsão realizada, devido ao acréscimo ou remoção de variáveis explicativas. Os 6 subconjuntos de atributos explicativos são:

- **Subconjunto 1**: contém 5 atributos, sendo estes, "Mes_Saida", "Dia_Semana_Saida", "Semana_Mes_Saida", "Feriado" e "Recesso_Escolar";
- **Subconjunto 2**: contém 7 atributos, sendo estes, "Ano_Saida", "Mes_Saida", "Dia_Saida", "Dia_Semana_Saida", "Semana_Mes_Saida", "Feriado" e "Recesso_Escolar";
- **Subconjunto 3**: contém 9 atributos, sendo estes, "Ano_Saida", "Mes_Saida", "Dia_Saida", "Dia_Semana_Saida", "Semana_Mes_Saida", "Vespera_Feriado", "Feriado", "Pos_Feriado" e "Recesso_Escolar";
- **Subconjunto 4**: contém 9 atributos, sendo estes, "Ano_Saida", "Mes_Saida",

"Dia_Saida", "Dia_Semana_Saida", "Semana_Mes_Saida", "Feriado", "Recesso_Escolar", "Preco_Gasolina" e "Preco_Passagem_Antecipada";

- **Subconjunto 5**: contém 10 atributos, sendo estes, "Ano_Saida", "Mes_Saida", "Dia_Saida", "Dia_Semana_Saida", "Semana_Mes_Saida", "Feriado", "Temperatura_Maxima", "Temperatura_Minima" e "Chuva";
- **Subconjunto 6**: contém 10 atributos, sendo estes, "Ano_Saida", "Mes_Saida", "Dia_Saida", "Dia_Semana_Saida", "Semana_Mes_Saida", "Feriado", "Recesso_Escolar", "Preco_Gasolina", "Preco_Passagem_Antecipada" e "Chuva".

Dado que 6 subconjuntos de atributos foram elencados para o treinamento e validação das redes, é trivial afirmar que existem 6 *"tipos"* de RNAs que foram construídas. Cada tipo possui o número p de nós da camada de entrada igual a quantidade de atributos que o subconjunto de dados utilizado para o treinamento de suas respectivas redes possui. Por exemplo, uma RNA treinada e validada com o Subconjunto 1 possui 5 nós na camada de entrada, enquanto que uma rede neural construída com base no Subconjunto 2 possui 7 nós de entrada, e assim sucessivamente. Em relação à camada de saída, todas as RNAs possuem apenas 1 neurônio, visto que existe apenas uma variável alvo (Quantidade de Passageiros por Dia).

Fica evidente, deste modo, que redes neurais treinadas e validadas segundo um mesmo subconjunto de atributos possuem o mesmo número de nós na camada de entrada e o mesmo número de neurônios na camada de saída. Contudo, o número de neurônios na camada oculta para redes do mesmo subconjunto foi determinado de acordo com a seguinte regra: a primeira RNA dentro de uma classe possui o número de neurônios ocultos igual à média entre o número de nós de entrada e o número de neurônios da camada de saída; a segunda RNA possui o número de neurônios da primeira RNA acrescido de 2; a terceira RNA possui o número de neurônios da segunda acrescido de 2 e a quarta RNA possui o número de neurônios da terceira acrescido de 2. Como resultado, para cada um dos 6 subconjuntos de atributos foram construídas 4 RNAs, o que totaliza em 24 redes neurais propostas.

A identificação das RNAs foi composta de acordo com a seguinte regra: começando com o prefixo "RNA_" para indicar que o modelo de previsão se trata de uma Rede Neural Artificial, depois acrescida do termo "Sn", onde n representa o número do subconjunto, seguido de um espaço em branco, e por último pela sequencia de caracteres "p-h-1" que representam respectivamente o número de nós de entrada, neurônios na camada oculta e neurônio de saída

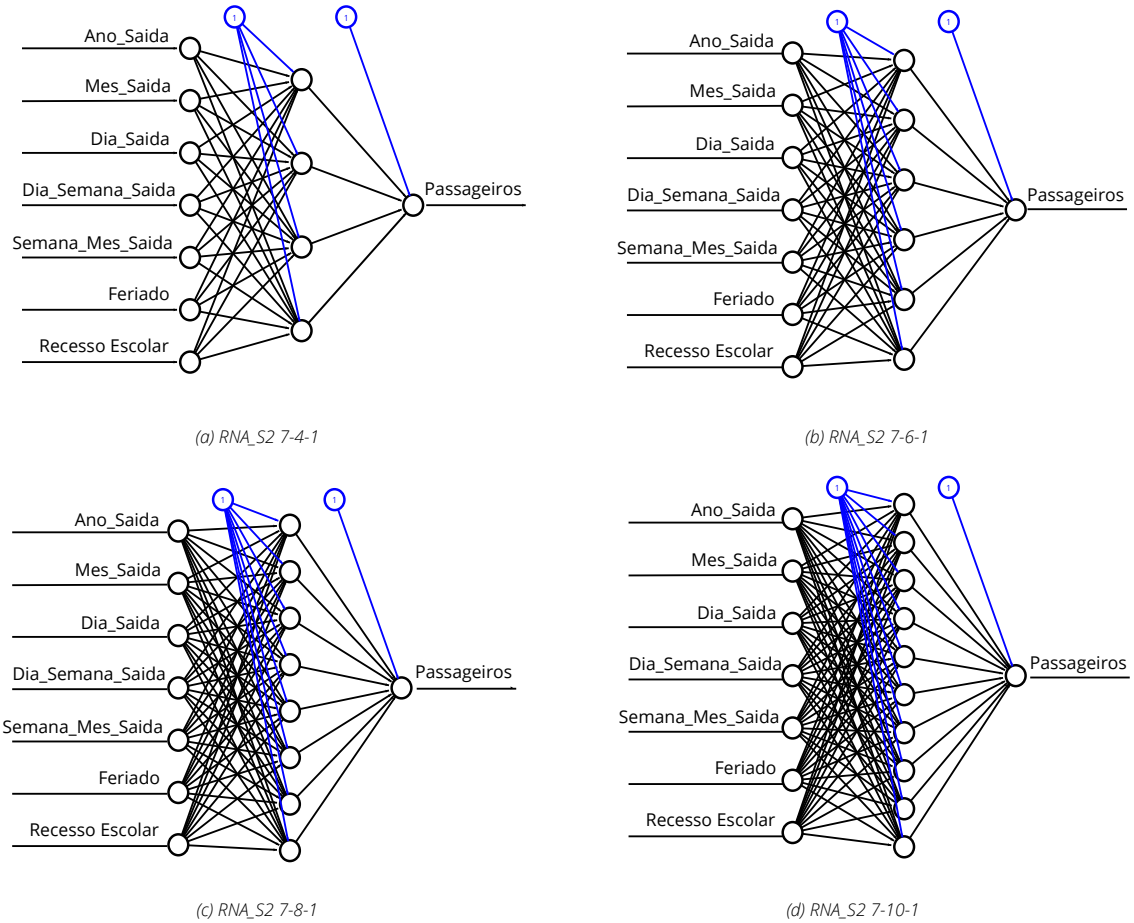
Para ilustrar o que foi descrito nos parágrafos anteriores, tome-se como exemplo as RNAs construídas para o Subconjunto 2 que possui um conjunto de dados com 7

atributos para treinamento e validação das redes. As RNAs construídas são:

- RNA_S2 7-4-1: 7 neurônios na camada de entrada, 4 na camada oculta e 1 na camadade de saída, Figura 21(a);
- RNA_S2 7-6-1: 7 neurônios na camada de entrada, 6 na camada oculta e 1 na camadade de saída, Figura 21(b);
- RNA_S2 7-8-1: 7 neurônios na camada de entrada, 8 na camada oculta e 1 na camadade de saída, Figura 21(c);
- RNA_S2 7-10-1: 7 neurônios na camada de entrada, 10 na camada oculta e 1 na camadade de saída, Figura 21(d).

Cada uma das 24 RNAs foram construídas utilizando-se a linguagem de programação R (R Core Team (2017)) e a Biblioteca “neuralnet”, (FRITSCH; GUENTHER, 2016). As RNAs são do tipo MLP, com uma camada oculta, *feedfoward*, completamente conectadas e utilizam função sigmoidal como função de ativação dos neurônios e o algoritmo *backpropagation* para ajuste dos pesos.

Figura 21 – RNAs Subconjunto 2.



(a) RNA_S2 7-4-1

(b) RNA_S2 7-6-1

(c) RNA_S2 7-8-1

(d) RNA_S2 7-10-1

Fonte: Autor, 2018.

3.6 Design Experimental

Esta seção tem por objetivo explicar a forma como se deu o processo de treinamento e validação de cada uma redes descritas na Seção 3.5.

A Biblioteca “neuralnet” (Fritsch e Guenther (2016)) permite a construção e treinamento de RNAs de forma relativamente simples. O usuário deve realizar a entrada da base de dados para o treinamento e indicar qual das colunas da base de dados representa a variável alvo e quais colunas representam os atributos explicativos, o número de camadas da rede, o número de neurônios por camada, o critério de parada do algoritmo de retropropagação de erro e o conjunto de pesos iniciais das sinapses. Uma opção é apresentada para caso o usuário queira iniciar o peso das sinapses de forma aleatória. Para o presente estudo todos os pesos de sinapses foram inicializados de forma aleatória.

Para o processo de validação a biblioteca “neuralnet” (Fritsch e Guenther (2016)) disponibiliza um algoritmo que, dada um rede neural já treinada e um conjunto de atributos explicativos, fornece a previsão da variável alvo para cada registro de atributos explicativos. Desta forma é possível comparar os valores preditos e os valores reais através, por exemplo, do Erro Médio Percentual Absoluto (EMPA) Equação (2.11b), que é utilizado como uma medida da qualidade do estimador.

De acordo com Seção 2.4.1, no processo de modelagem de uma rede neural, uma base de dados deve ser dividida em uma parte para treinamento e outra parte para validação da rede. Não existem diretrizes para a melhor proporção treinamento-validação, por isso, decidiu-se que 3 proporções seriam avaliadas para cada um dos subconjuntos de dados apresentados na Seção 3.5. Sendo a primeira proporção de 70% dos dados para treinamento e os 30% restantes para validação, a segunda, de 80% para treinamento e 20% para validação, e a terceira, de 90% para treinamento e 10% para validação.

O processo “treinamento-validação” ocorre do seguinte modo: começando pelo Sunconjunto 1, cada uma das 4 redes pertencentes ao subconjunto em questão recebe 70% dos registros para realizar o treinamento. Após o término do treinamento, as redes recebem os 30% dos dados restantes sem a variável alvo, para que, baseado nos atributos explicativos, as redes façam a previsão da variável alvo. Depois de realizada as previsões, para cada uma das 4 redes dentro do Subconjunto1 é calculado o valor do EMPA, que por sua vez, é armazenado em um vetor para uso posterior. A esta sequência deu-se o nome de rodada.

Como o precesso de treinamento é dependente dos pesos iniciais das sinapses, e, variando-se os pesos iniciais pode-se chegar a valores diferentes de EMPA para a mesma arquitetura de rede, decidiu-se por realizar 30 rodadas de “treinamento-validação” para cada uma das 4 redes presentes no Subconjunto 1. Sendo assim, cada

rede da classe RNA_S1 possui um vetor correspondente com 30 valores de EMPA para a proporção treinamento-validação em questão. Tendo uma amostra de 30 EMPA é possível calcular a média dos EMPA bem como o desvio padrão da amostra para cada RNA, como será apresentado no Capítulo 4.

Após as 30 rodadas de treinamento-validação, para cada rede dentro do Subconjunto 1 utilizando a proporção $70 - 30$, é realizado o mesmo número de rodadas para as proporções $80 - 20$ e $90 - 10$. Esgotadas as proporções, um novo subconjunto é analisado do mesmo modo que o Subconjunto 1. A Figura 22 apresenta um pseudocódigo do processo descrito anteriormente.

Figura 22 – Pseudocódigo processo aprendizagem-validação.

Pseudocódigo Processo de avaliação das RNAs

Entradas:

S - conjunto de sunconjunto de atributos

P - conjunto de proporções

n - número de vezes que será realizado o processo de treino-validação para cada RNA dos subconjuntos

Saída:

Média e desvio padrão do EMPA para n rodadas de todas RNAs, dada uma proporção e conjunto de atributos explicativos

```
1  Início
2      Vetor Resultado = [ ]
3      para cada subconjunto em S faça:
4          para cada proporção em P faça:
5              reparta o subconjunto em proporção de Treino e proporção de Teste
6              para cada RNA i presente no subconjunto:
7                  Vetor de n EMPA = [ ]
8                  faça n vezes:
9                      Aprendizagem com a proporção de Treino
10                     Validação com a proporção de Teste
11                     Cálculo do EMPA do n-ésimo Teste
12                     Armazene EMPA n-ésimo Teste em Vetor n EMPA
13             fim
14             Calcule a média e desvio padrão de n valores de EMPA para a RNA i do subconjunto a uma dada proporção
               utilizando o Vetor de n EMPA
15             Armazene a média e o desvio padrão para a RNA i do subconjunto a uma dada proporção no Vetor Resultado
16         fim
17     fim
18     Retorne Vetor Resultado
19 fim
```

Fonte: Autor, 2018.

Também foi proposto que, para a díade “subconjunto de atributos explicativos”/“arquitetura da rede” que apresentasse o menor EMPA, fosse realizado mais uma rodada de treinamento-validação, entretanto, aplicando-se o conceito de janela deslizante. O conceito de janela deslizante pode ser entendido como o processo no qual é utilizado um intervalo de dados de um tamanho pré-determinado (janela) para o treinamento da rede e os dados preditos são os registros subsenquentes do final da janela até um tamanho de passo. À medida que a rede conclui a etapa de predição, a

janela é movimentada de acordo com o tamanho de passo, o que resulta na exclusão de um intervalo de dados do tamanho de passo do início da janela e na adição de um intervalo de dados do tamanho de passo no final. Por exemplo, escolhendo-se o tamanho da janela igual a 450 registros e o tamanho de passo igual a 7 registros, o primeiro "treino" da rede aconteceria utilizando-se os registros de 1 a 450 e seriam previstos os registros 451, 452, 453, 454, 455, 456 e 457. Na próxima iteração a janela seria deslocada em sete unidades, logo o intervalo de treinamento seria do registro 8 a 457 e seriam previstos os registros 458, 459, ..., 464, e assim sucessivamente até que a base de dados chegasse ao fim.

É importante apontar que cada vez que a janela se movimenta um novo processo de treinamento se inicia no qual é mantida a arquitetura da rede, mas não os pesos iniciais das sinapses (que são reiniciados aleatóriamente). A Seção 4.2 apresenta os resultados obtidos com o processo de janela deslizante.

4 RESULTADOS

Este Capítulo visa apresentar os resultados obtidos através dos processos de treinamento e validação das RNAs descritas na Seção 3.6. A Seção 4.1 apresenta os resultados das RNAs construídas com o método de proporção para treinamento-validação, ao passo que a Seção 4.2 mostra os resultados da RNA treinada e validada pelo processo de janela deslizante.

4.1 Proporção treinamento-validação

Dado que o objetivo geral deste trabalho é avaliar a aplicação de RNAs como modelos de previsão de passageiros no transporte público, primeiramente fez-se necessário o cálculo da *Taxa Média de Ocupação Diária* da Linha 0700. Este valor fornece uma estimativa da disparidade entre a oferta e a demanda de lugares no transporte coletivo para o período de dados analilsados, servindo como um indicador da oportunidade *teórica* máxima de adequação da relação oferta-demanda.

O termo *teórica* é utilizado pelo fato de que operacionalmente seria pouco provável o *ajuste perfeito* entre oferta e demanda por, principalmente, dois motivos. Primeiro, existem casos em que, para algumas faixas horárias ao longo do dia, a quantidade de passageiros é consideravelmente menor do que a capacidade do veículo utilizado para a operação das linhas de transporte nos *dois* sentidos (quando se tratam de linhas troncais e inter-terminais). Neste caso, a disparidade entre oferta e demanda, poderia ser resolvida de duas formas: pela aquisição de veículos de menor capacidade para operar dentro dessas faixas horárias ou pelo aumento do *headway* (tempo entre partidas de viagens) visando o aumento do número de passageiros partindo da origem. Contudo, ambas as soluções devem ser utilizadas com parcimônia porque o aumento demasiado da frota pode resultar no aumento do custo de manutenção e armazenamento de veículos, e o aumento do *headway* resulta no aumento do tempo de viagem do usuário, o que pode tornar o modal menos atrativo.

O segundo motivo impeditivo para a adequação ideal entre a oferta e a demanda são os casos em que as cidades possuem divisões claras entre as regiões com a maior parte dos empregos e as regiões com a maior parte da concentração habitacional. Este tipo de organização do uso do solo resulta no acúmulo de viagens em um único sentido durante os períodos de pico de tráfego.

No caso específico da Cidade de Joinville, para o pico da manhã, por exemplo,

o número de viagens no sentido Bairro-Centro são maiores do que o número de viagens no sentido Centro-Bairro. Sendo assim, um veículo com capacidade adequada para realizar determinada linha de ônibus no sentido Bairro-Centro no pico da manhã, estará superdimensionado para realizar o sentido contrário da mesma linha no mesmo período. O que resultará em uma taxa de ocupação média relativamente baixa se forem considerados os números de passageiros e lugares ofertados dos dois sentidos de viagem. Algumas soluções poderiam ser tomadas, a longo prazo, para a diminuição do acúmulo de viagens em um sentido durante o pico. Dentre estas estão: modificações nas Leis de Uso do Solo para incentivar o desenvolvimento econômico, e com este, a geração de empregos, em todas as áreas da cidade, e o incentivo à diferenciação no horário de início e fim de jornada de trabalho e estudo, entre empresas e instituições educacionais para que os períodos de picos fossem diluídos ao longo do dia.

Apesar de não indicar uma oportunidade de ganho real, do ponto de vista operacional, pelos motivos discutidos anteriormente, a *Taxa Média de Ocupação Diária* é, matematicamente, um indicador que caracteriza a discrepância entre oferta e demanda de transporte e foi utilizada no presente estudo como referência para avaliar os ganhos provenientes da utilização de RNAs como previsores da demanda de passageiros.

A base de dados do transporte coletivo, apresentada na Seção 3.2.1, contém para cada registro de viagem o atributo “Identificador do Veículo”. Desta forma, através dos dados a respeito da capacidade de cada veículo (visto que a capacidade do ônibus representa o número de lugares ofertados para uma viagem) foi possível calcular a “Quantidade de Lugares Ofertados por Dia”. Somando-se todos os valores para a variável “Quantidade de Lugares Ofertados por Dia” (Equação (4.1a)) e, para a variável “Quantidade de Passageiros por Dia” (Equação (4.1b)), obtém-se, respectivamente, a quantidade total de lugares ofertados, bem como a quantidade total de passageiros para o período. A Taxa Média de Ocupação Diária é o resultado da divisão entre o total de passageiros pelo total de lugares, Equação (4.1c).

$$QTLO = \sum_{n} QLOD_i \qquad \forall\, i = 1, 2, ..., n. \tag{4.1a}$$

$$QTP = \sum_{n} QPD_i \qquad \forall\, i = 1, 2, ..., n. \tag{4.1b}$$

$$TMOD = \frac{QTP}{QTLO} \tag{4.1c}$$

Onde:

n - Número de dias dentro do período analisado;

$QTLO$ - Quantidade total de lugares ofertados para o período;

$QLOD$ - Quantidade de lugares ofertados por dia;

QTP - Quantidade total de passageiros para o período;
QPD - Quantidade de passageiros por dia;
$TMOD$ - Taxa média de ocupação diária.

Para a Linha 0700, no período de $01/05/2015$ a $30/04/2017$, a Taxa Média de Ocupação Diária é de aproximadamente $46,28\%$, ou seja, a Linha opera com $53,72\%$ de sua capacidade em ociosidade.

Para efeitos comparativos foi proposto um estimador secundário da variável alvo “Quantidade de Passageiros por Dia” baseado no atributo “Dia_Semana_Saída”. O Estimador em questão baseia-se no fato de que o *Preditor base* para uma variável qualquer é a média dos valores da variável analisada. Entretanto, é sabido que existem variações consideráveis na quantidade de passageiros por dia ao longo da semana, principalmente entre dias úteis e finais de semana. Por esse motivo, ao invés de utilizar o valor da média da “Quantidade de Passageiros por Dia” de todos os 731 registros presentes na “Base Geral de Dados”, foi proposto calcular a média da “Quantidade de Passageiros por Dia” para cada dia da semana: *Domingo*, *Segunda*, *Terça*, ..., *Sábado*.

O processo de treinamento-validação do Preditor baseado nos dias da semana ocorre da seguinte maneira: a “Base Geral de Dados” é dividida na proporção de 90% dos registros para treino e 10% dos registro para validação. Com a proporção de treino são calculadas as médias da quantidade de passageiros por dia da semana. Na amostra de validação, o atributo “Dia_Semana_Saída” é usado como *entrada* e então, como valor predito para a variável alvo, é atribuído o valor da quantidade média de passageiros por dia da semana. O EMPA obtido com o Preditor por dia da semana foi de $17,58\%$. Ou seja, hipoteticamente, se o *Preditor base* fosse utilizado para a determinação da oferta, a discrepância entre a oferta e a demanda caíria de $53,72\%$ para $17,58\%$.

As Figuras 23, 24, 25, 26, 27 e 28 apresentam o valor do EMPA para cada proporção de treinamento-validação e para cada configuração de neurônios na camada oculta dos Subconjunto 1, Subconjunto 2, Subconjunto 3, Subconjunto 4, Subconjunto 5 e Subconjunto 6, respectivamente. Nas figuras citadas anteriormente, destaque na cor laranja é dado para a menor média de EMPA. Esse valor destacado representa o *EMPA do melhor modelo* dentro dos subconjuntos.

Para todos os Subconjuntos, os valores de EMPA do melhor modelo ocorreram para a proporção de treinamento-validação de $90-10$, o que pode indicar que quanto maior a base de dados para treino, melhor é o ajuste dos pesos sinápticos devido ao fato de que mais registros referentes a um mesmo período do ano, em anos diferentes, podem ter sido utilizados para treinamento. Por exemplo, dado que o período dos dados se extendem de $01/05/2015$ a $30/04/2017$, na proporção $70-30$ existe apenas um registro referente ao dia de Natal ($25/12/2015$) na fatia de dados usada para treino

(70%), ao passo que, na fatia de treino da proporção $90 - 10$ existem dois registros para o dia de Natal ($25/12/2015$ e $25/12/2016$).

Outra característica observada entre os subconjuntos foi que os valores de EMPA do melhor modelo ocorreram para as RNAs com arquitetura em que o número de neurônios na camada oculta era igual a média entre o número de neurônios da camada de entrada e da camada de saída.

O Subconjunto 1 apresenta a maior média de EMPA do melhor modelo, com um valor de $21,29\%$, o que faz com que esse subconjunto gere um previsor da "Quantidade de Passageiros ao longo do Dia" pior que o *Estimador base* que possui valor de EMPA igual a, $17,58\%$. Uma possível explicação para este resultado é que a capacidade de aprendizagem das RNAs que utilizam o Subconjunto 1 foi comprometida pelo fato de existerem apenas atributos explicativos que expressam as caracteristiscas sazonais da demanda de passageiros, sendo estes o "Mes_Saida", "Dia_Semana_Saída", "Semana_Mes_Saida", "Recesso_Escolar" e um atributo explicativo da componente aleatória da demanda expressa pela variável "Feriado". De acordo com a Seção 3.4 pode-se observar que a "Quantidade de Passageiros ao longo do Dia" possui uma componente de tendência expressiva.

Figura 23 – Resumo estatístico das redes neurais do Subconjunto 1.

Subconjunto de dados 1 - Mes_Saida; Dia_Semana_Saida; Semana_Mes_Saida; Feriado; Recesso_Escolar.							
ID Rede	Neur×nios Camada Oculta	70% da amostra para treino		80% da amostra para treino		90% da amostra para treino	
		M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA
RNA_S1 5-3-1	3	23.782	1.649	26.890	3.001	21.294	0.562
RNA_S1 5-5-1	5	23.362	0.940	24.783	1.405	22.074	1.585
RNA_S1 5-7-1	7	23.037	0.939	24.499	1.322	23.255	1.638
RNA_S1 5-9-1	9	23.006	0.678	24.081	0.887	24.574	1.410

Fonte: Autor, 2018.

Em oposição ao Subconjunto 1, o Subconjunto 2 possui a menor média, dentre todos os subconjuntos, de EMPA do melhor modelo, sendo este valor de 11.97%; uma melhora considerável em relação ao EMPA do Subconjunto 1 bem como do *Estimador base*. Esta melhora no valor do EMPA pode ser resultado da adição de atributos que expressam a componente de tendência da demanda de passageiros, pois o Subconjunto 2 possui os mesmos atributos explicativos do Subconjunto 1 acrescido dos atributos "Ano_Saída" e "Dia_Saída".

O Subconjunto 3 é basicamente o Subconjunto 2 mais os atributos "Vespera_Feriado" e "Pos_Feriado". A decisão de adicionar esses atributos na "Base Geral de Dados" se deu pelo motivo de que quando um feriado qualquer acontece na Terça-feira ou na Quinta-feira, a "Quantidade de Passageiros ao Longo do Dia" do dia anterior, no caso da Terça-feira, e do dia posterior no caso da Quinta-feira, podem

Figura 24 – Resumo estatístico das redes neurais do Subconjunto 2.

Subconjunto de dados 2 - Ano_Saida; Mes_Saida; Dia_Saida; Dia_Semana_Saida; Semana_Mes_Saida; Feriado; Recesso_Escolar.							
ID Rede	Neur×nios Camada Oculta	70% da amostra para treino		80% da amostra para treino		90% da amostra para treino	
		M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA
RNA_S2 7-4-1	ぬ	ねねんん	ぅ゛	わぬんゎ	わろろん	ろぅんわ	ろ愛ね
RNA_S2 7-6-1	を	わうねわ゛	ろねめろ	わうねぬ	ねいをを゛	ろわんされ	わをれ
RNA_S2 7-8-1	う	わうねされ	ろろんをん	わんわゑ	ろうわうろ	ろぬんれ	わねをを
RNA_S2 7-10-1	ろれ	わろうわ	ろわ愛を	ぬわ愛を	ろをめろ	ろうろぬ	ゑねろわ

Fonte: Autor, 2018.

sofrer alteração. Contudo, para o período de dados analisados, a adição dos atributos “Vespera_Feriado” e “Pos_Feriado” nos dados de treino não se traduziu em melhora no processo de treinamento, pois a média de EMPA foi de $14,36\%$.

Figura 25 – Resumo estatístico das redes neurais do Subconjunto 3.

Subconjunto de dados 3 - Ano_Saida; Mes_Saida; Dia_Saida; Dia_Semana_Saida; Semana_Mes_Saida; Vespera_Feriado; Feriado; Pos_Feriado; Recesso_Escolar.							
ID Rede	Neur×nios Camada Oculta	70% da amostra para treino		80% da amostra para treino		90% da amostra para treino	
		M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA
RNA_S3 9-5-1	ゑ	わをねろ	ゐぬ	わんんわ	ろわろろ	ろぬわ愛ぅ	わねぅ
RNA_S3 9-7-1	ん	わんねぬ゛	ろろねわ	わぅんゑ	ろわ愛ぅ゛	ろをねねわ	わろを
RNA_S3 9-9-1	゛	わろをを	ぅをぬ	わ愛ねぅ	ろうんわ	ろぅねぬ	ぬ愛ゑ
RNA_S3 9-11-1	ろろ	わ ぬぬ	ろねねど	わんねろ	ろぬろぬ	ろぅねぅ	ゑねわ

Fonte: Autor, 2018.

Os Sunconjuntos 4, 5 e 6 são variações do Subconjunto 2 no que diz respeito a adição de atributos explicativos econômicos, climáticos e ambos, respectivamente. Os valores de média de EMPA encontrados para estes Subconjuntos foi da ordem de 12%.

Figura 26 – Resumo estatístico das redes neurais do Subconjunto 4.

Subconjunto de dados 4 - Ano_Saida; Mes_Saida; Dia_Saida; Dia_Semana_Saida; Semana_Mes_Saida; Feriado; Recesso_Escolar; Preco_Gasolina; Preco_Passagem_Antecipada.							
ID Rede	Neur×nios Camada Oculta	70% da amostra para treino		80% da amostra para treino		90% da amostra para treino	
		M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA
RNA_S4 9-5-1	ゑ	わゑろを	をねねゑ	わんろねゑ	をねわれ	ろねわぅど	ろぬわ
RNA_S4 9-7-1	ん	わろねをわ	をねねゑ	わ れいわ	ろわねを	ろわをろん	ろんわろ
RNA_S4 9-9-1	゛	わわをろ゛	をぬん゛	わゑんれ	んわわゑ	ろわろど	ろねを
RNA_S4 9-11-1	ろろ	わろをゑ゛	をぬねろ	わうねど	ろわをぬ	ろわをぬ	わねゑゑ

Fonte: Autor, 2018.

Os Subconjuntos 2, 4, 5 e 6 possuem valores próximos entre si de média de

Figura 27 – Resumo estatístico das redes neurais do Subconjunto 5.

Subconjunto de dados 5 - Ano_Saida; Mes_Saida; Dia_Saida; Dia_Semana_Saida; Semana_Mes_Saida; Feriado; Recesso_Escolar; Temperatura_Maxima; Temperatura_Minima; Chuva.							
ID Rede	Neur×nios Camada Oculta	70% da amostra para treino		80% da amostra para treino		90% da amostra para treino	
		M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA
RNA_S5 10-6-1	6	26.675	10.232	37.846	19.754	12.717	1.909
RNA_S5 10-8-1	8	34.527	12.854	40.183	21.040	14.349	2.399
RNA_S5 10-10-1	10	28.658	9.917	43.250	16.129	15.195	3.317
RNA_S5 10-12-1	12	35.777	9.857	45.498	16.767	16.521	6.129

Fonte: Autor, 2018.

Figura 28 – Resumo estatístico das redes neurais do Subconjunto 6.

Subconjunto de dados 6 - Ano_Saida; Mes_Saida; Dia_Saida; Dia_Semana_Saida; Semana_Mes_Saida; Feriado; Recesso_Escolar; Preco_Gasolina; Preco_Passagem_Antecipada; Chuva.							
ID Rede	Neur×nios Camada Oculta	70% da amostra para treino		80% da amostra para treino		90% da amostra para treino	
		M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA	M¶dia EMPA	Desvio Padr² o EMPA
RNA_S6 10-6-1	を	わ					
RNA_S6 10-8-1	う						
RNA_S6 10-10-1	れ						
RNA_S6 10-12-1	ろ						

Fonte: Autor, 2018.

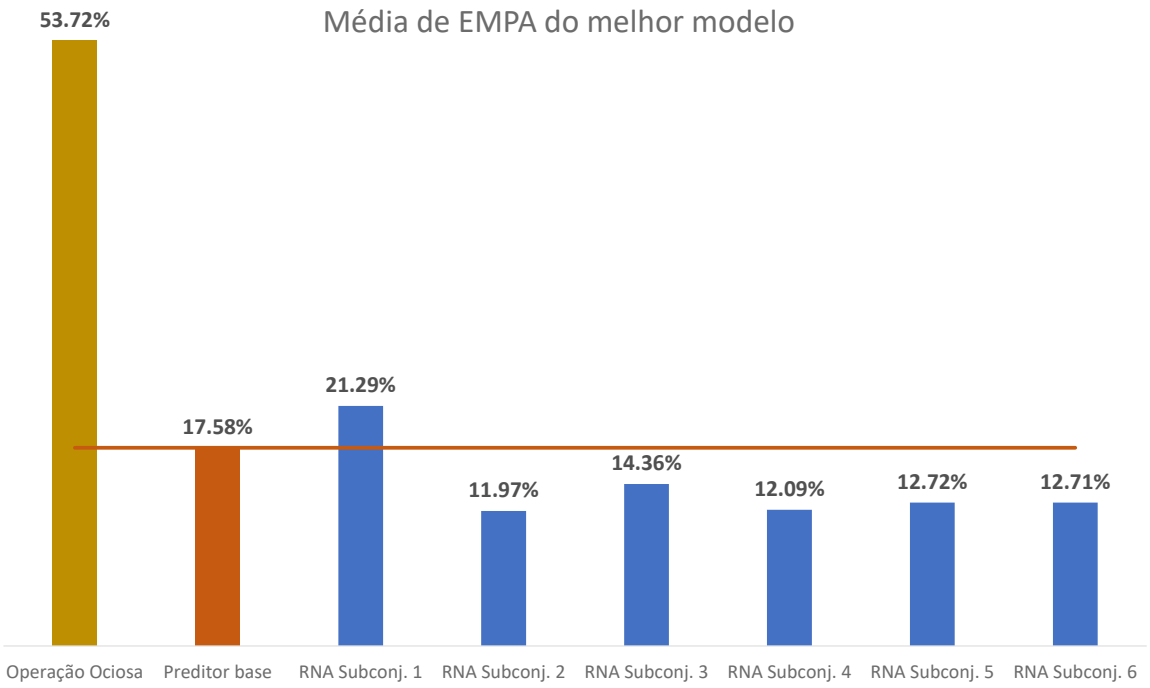
EMPA do melhor modelo, Figura 29. Tendo em vista que, o custo da etapa de pré-processamento é diretamente proporcional à quantidade de atributos de uma base de dados, o Subconjunto 2 demonstra o melhor custo-benefício relativo em termos de pré-processamento dos dados e valor de EMPA.

Por possuir o menor EMPA dentre todos os subconjuntos e todas as arquiteturas, o Subconjunto de atributos explicativos 2 e a Rede RNA_S2 $7-4-1$ foi utilizada para construção de uma rede com janela deslizante apresentada na Seção 4.2.

4.2 Rede Neural com janela deslizante

De acordo com a Seção 3.6 a Rede Neural RNA_S2 $7-4-1$ foi utilizada para a construção de uma Rede Neural com Janela deslizante de tamanho de 450 dados, aproximadamente 15 meses, e tamanho de passo com 7 dados, o que corresponde a uma semana. Ou seja, o método de janela deslizante realiza a aprendizagem baseado em 15 meses de dados e prediz os valores da semana seguinte. Depois, a janela é deslocada em uma semana, treina-se em mais 15 meses de dados e se prediz a próxima semana e assim sucessivamente até que se esgote a base de dados.

Figura 29 – Comparação da média de EMPA entre as melhores RNAs utilizadas.

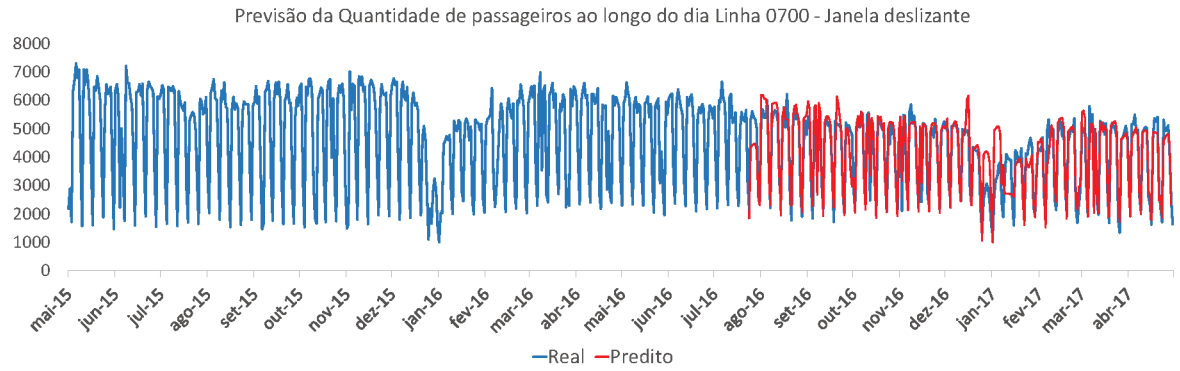


Fonte: Autor, 2018.

O valor do EMPA para a Rede RNA_S2 $7-4-1$ com Janela deslizante, foi de $11,72\%$. Resultado discretamente melhor se comparado aos 11.97% obtido para a mesma configuração de rede pelo processo de proporção aprendizagem-validação. A Figura 30 apresenta o gráfico que retrata a diferença entre os valores reais do número de passageiros por dia (curva de cor azul) e os valores preditos (curva de cor vermelha). É importante observar que na Figura 30 não existem previsões para os primeiros 450 dados, pois estes dados fazem parte da amostra inicial de treinamento.

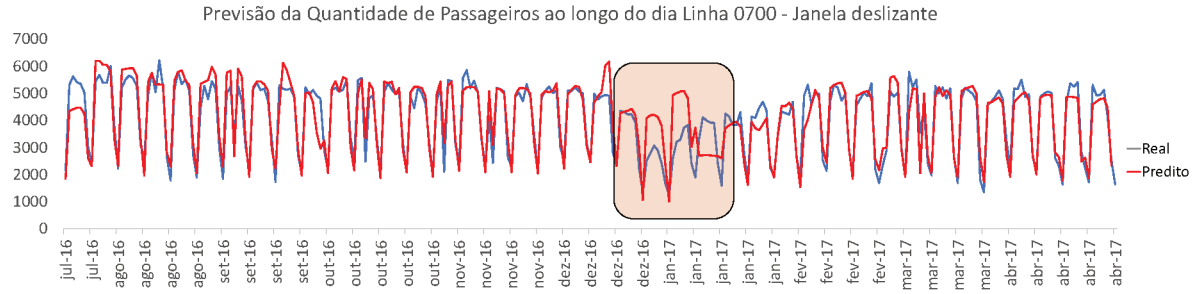
A Figura 31 apresenta os mesmo dados retratados na Figura 30, entretanto, somente para o período em que os dados começam a ser preditos, isto é, a partir do registro 451 da “Base Geral de Dados”. Destaque foi dado para o período que se estende da última semana do mês de Dezembro de 2016 até a segunda semana do mês de Janeiro de 2017 que é o período no qual a Rede Neural realiza as predições com maior valor de EMPA unitário.

Figura 30 – Comparação entre os valores reais de passageiros por dia e os valores preditos pela Janela Deslizante.



Fonte: Autor, 2018.

Figura 31 – Destaque dos valores preditos com maior EMPA.



Fonte: Autor, 2018.

Uma possível explicação para a ocorrência de previsões com EMPA relativamente alto, é que a última semana de Dezembro e a primeira semana de Janeiro, apesar de serem classificadas como um período de férias escolares, foram tratadas como semanas normais de funcionamento do comércio e empresas o que não ocorre na prática de acordo com a “Base Geral de Dados”. Outro fator que contribui para a presença de previsões tão discrepantes é o tamanho da janela de treino ser relativamente pequena, pois para 450 registros (15 meses de dados) existem vários dias que só aparecem uma única vez na amostra de treino. Um tamanho de Janela mais justo seriam 2 anos de dados, o que é igual ao tamanho da “Base Geral de Dados”. Sendo assim, para melhores resultados com Janelas Deslizantes seriam necessários bases de dados maiores.

5 CONCLUSÕES

O adequado atendimento da demanda por lugares no transporte coletivo se apresenta como uma forma que as operadoras de transporte público possuem de, simultaneamente, fornecer o nível de serviço estipulado pelo governo municipal e realizar a operação praticando custos menores. Contudo, a demanda por viagens se torna uma variável difícil de ser predita, primeiramente, devido a sua multidimensionalidade, e segundo, ao alto custo associado com métodos de pesquisas tradicionais para levantamento de dados.

À medida que Sistemas Inteligentes de Transporte, como é o caso da bilhetagem eletrônica e do GPS, se tornam mais difundidos, o processo de aquisição de dados se torna automatizado e com isso têm-se acesso a volumes de dados maiores e com mais qualidade, o que por sua vez, acarreta na necessidade da utilização de métodos não tradicionais de análise, como é o caso dos métodos da Área de Mineração de Dados que engloba métodos de análise estatística e algoritmos da área de Aprendizagem de Máquina.

Nesse contexto, para a construção de modelos a partir de bases de dados, Redes Neurais Artificiais têm se mostrado uma ferramenta robusta para o tratamento de dados, principalmente em problemas que envolvem a predição de uma variável de interesse baseada nos valores de outras variáveis explicativas para o fenômeno.

O presente trabalho utilizou RNAs para a previsão da quantidade de passageiros por dia em uma linha do transporte público da cidade de Joinville que opera com aproximadamente 54% de sua capacidade em ociosidade. Para a linha em questão as RNAs construídas geraram um Erro Médio Percentual Absoluto entre a quantidade real de passageiros e a quantidade predita da ordem de 11%, o que mostra o potencial do uso de redes neurais como ferramentas auxiliadoras do planejamento operacional de transportes. É importante considerar que, como o EMPA representa o erro absoluto, a diferença entre a previsão e o valor observado pode ser tanto para mais como para menos. Caso a previsão indique uma demanda de passageiros maior do que a demanda real, o “desperdício” de recurso será convertido imediatamente em conforto para o usuário, entretanto, previsões abaixo do valor real podem resultar em superlotação.

Muitos desafios ainda precisam ser superados na área de previsão de demanda utilizando RNAs. Como sugestões de trabalhos futuros, para tornar a previsão do número de passageiros com maior utilização operacional, como por exemplo, na

determinação da frequência de viagens e capacidade de veículo utilizado para viagens, sugere-se a redução do período de previsão de passageiros por dia para passageiros por faixa horária. Trabalhos também poderiam ser realizados com a intenção de aumentar a “Base Geral de Dados” para que se possa utilizar RNAs com janelas deslizantes maiores. Ainda, um modelo de produção de viagens na origem utilizando RNAs poderia ser utilizado para a construção de um modelo gravitacional unicamente restrito de Distribuição de Viagens, o que resultaria em uma Matriz OD por ônibus para a área de estudo.

REFERÊNCIAS

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS. **Série histórica do levantamento de preços e de margens de comercialização de combustíveis. Relatório semanal de 2013 a 2018**. [S.l.], 2018. Disponível em: <http://www.anp.gov.br/precos-e-defesa/234-precos/levantamento-de-precos/>.

ARBIB, M. A. **Brains, Machines, and Mathematics**. 2. ed. [S.l.]: Springer-Verlag, New York., 1987.

ASSOCIAÇÃO NACIONAL DAS EMPRESAS DE TRANSPORTES URBANOS. **Transporte de passageiros por modo Rodoviário 2018**. Brasília, 2017. Disponível em: <http://www.ntu.org.br/novo/AreasInternas.aspx?idArea=7>.

BRASIL. Constituição da República Federativa do Brasil de 1988. **Diário Oficial da União** — Seção 1, Brasilília, DF., p. 1–32, 1988.

BRASIL. Lei nº 12.587, de 3 de janeiro de 2012 Institui as diretrizes da Política Nacional de Mobilidade Urbana. **Diário Oficial da União** — Seção 1, Brasilília, DF., p. 1–3, 2012.

BRASIL. **Política Nacional de Mobilidade Urbana**. Brasil, 2013. Disponível em: <http://www.portalfederativo.gov.br/noticias/destaques/municipios-devem-implantar-planos-locais-de-mobilidade-urbana/CartilhaLei12587site.pdf>.

BRASIL. **Feriados e Pontos Facultativos Oficiais do Brasil para 2015**. [S.l.], 2015. Disponível em: <http://www.brasil.gov.br/governo/2015/02/pais-tera-nove-feriados-e-sete-pontos-facultativos-em-2015>.

BRASIL. **Feriados e Pontos Facultativos Oficiais do Brasil para 2016**. [S.l.], 2016. Disponível em: <http://www.planejamento.gov.br/assuntos/gestao-publica/noticias/divulgado-calendario-de-feriados-nacionais-de-2016>.

BRASIL. **Feriados e Pontos Facultativos Oficiais do Brasil para 2017**. [S.l.], 2017. Disponível em: <http://www.brasil.gov.br/governo/2016/11/confira-feriados-nacionais-e-pontos-facultativos-de-2017>.

CAMPOS, V. B. G. **Planejamento de transportes. Conceitos e modelos**. 1. ed. [S.l.]: Interciência Rio de Janeiro, 2013.

CASEY, H. J. Applications to traffic engineering of the law of retail gravitation. **Traffic Quarterly**, v. 9, p. 23 – 25, 1955.

DEPARTMENT OF INFRASTRUCTURE AND REGIONAL DEVELOPMENT. **Trends. Infrastructure and transport to 2030**. Austrália, 2014. Disponível em: <https://infrastructure.gov.au/infrastructure/publications/files/Trends_Infrastructure_and_Transport_to_2030.pdf>.

DRAGANA, M.; MILICA Šelmić et al. Neural network based model for predictiong the number of sleeping cars in rail transport. **International Journal for Traffic and Transport Engineering**, p. 29 – 35, 2015.

EUROPEAN COMISSION. **EU energy, transport and GHG emissions. Trends to 2050**. European Union, 2013. Disponível em: <http://ec.europa.eu/transport/sites/transport/files/media/publications/doc/trends-to-2050-update-2013.pdf>.

EVANS, J. R.; LINDSAY, W. M. **Managing for quality and performance excellence**. 10. ed. [S.l.]: Cengage learning, 2014.

FERRAZ, A. C. P.; TORRES, I. G. E. **Tranporte Público Urbano**. 2. ed. [S.l.]: Rima São Carlos, 2004.

FOELL, S.; PHITHAKKITNUKOON, S. et al. Predictability of public transport usage: a study of bus rides in lisbon, portugal. **IEEE Transactions on Intelligent Tranportations Systems**, v. 16, p. 2955 – 2960, 2015.

FRITSCH, S.; GUENTHER, F. **neuralnet: Training of Neural Networks**. [S.l.], 2016. R package version 1.33. Disponível em: <https://CRAN.R-project.org/package=neuralnet>.

FUNDAÇÃO INSTITUTO DE PESQUISA E PLANEJAMENTO PARA O DESENVOLVIMENTO SUSTENTÁVEL DE JOINVILLE. **Cidade em Dados 2011**. Joinville - SC, 2011. Disponível em: <https://www.joinville.sc.gov.br/wp-content/uploads/2016/01/joinville-cidade-em-dados-2010-2011.pdf>.

HAYKIN, S. **Neural networks and learning machines**. 3. ed. [S.l.]: Pearson Education, Inc. New Jersey, 2009.

HILLIER, F. S.; LIEBERMAN, G. J. **Introdução à pesquisa operacional**. 9. ed. [S.l.]: AMGH Porto Alegre, 2013.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **População Município de Joinville 2017**. Rio de Janeiro, 2018. Disponível em: <https://cidades.ibge.gov.br/brasil/sc/joinville/panorama>.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **População Município de São Paulo 2017**. Rio de Janeiro, 2018. Disponível em: <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama>.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **População Município do Rio de Janeiro 2017**. Rio de Janeiro, 2018. Disponível em: <https://cidades.ibge.gov.br/brasil/rj/rio-de-janeiro/panorama>.

JOINVILLE. **Decreto Nº 23.592 de 29 de Dezembro de 2014. Reajuste do Preço da Passagem do Transporte Público Coletivo.** [S.l.], 2014. Disponível em: <https://leismunicipais.com.br/a/sc/j/joinville/decreto/2014/2360/23592/decreto-n-23592-2014/>.

JOINVILLE. **Calendário letivo do Município de Joinville 2015**. [S.l.], 2015. Disponível em: <https://www.joinville.sc.gov.br/?s=Calend%C3%A1rio+Escolar+2015>.

JOINVILLE. **Decreto Nº 26.192 de 28 de Dezembro de 2015: Reajuste do Preço da Passagem do Transporte Público Coletivo.** [S.l.], 2015. Disponível em: <https://leismunicipais.com.br/a2/sc/j/joinville/decreto/2015/2620/26192/decreto-n-26192-2015/>.

JOINVILLE. **Calendário letivo do Município de Joinville 2016**. [S.l.], 2016. Disponível em: <https://www.joinville.sc.gov.br/publicacoes/calendario-escolar-municipal-2016/>.

JOINVILLE. **Calendário letivo do Município de Joinville 2017**. [S.l.], 2017. Disponível em: <https://www.joinville.sc.gov.br/publicacoes/calendario-escolar-2017/>.

JOINVILLE. **Dados da Rede de Monitoramento das Estações Meteorológicas de Joinville/SC.** [S.l.], 2017. Disponível em: <https://prefeituradigital.joinville.sc.gov.br/servico/detalhe-61-Rede+de+Monitoramento.html>.

JOINVILLE. **Decreto Nº 28.169 de 03 de Janeiro de 2017: Reajuste do Preço da Passagem do Transporte Público Coletivo.** [S.l.], 2017. Disponível em: <https://leismunicipais.com.br/a/sc/j/joinville/decreto/2017/2817/28169/decreto-n-28169-2017/>.

JOINVILLE. **Histórico Festivais de Dança de Joinville**. [S.l.], 2017. Disponível em: <http://festivaldedancadejoinville.com.br/historico/>.

LONGLEY, P. A. L. et al. **Sistemas e ciênciencia da informação geográfica**. [S.l.]: Bookman, 2013.

EL MAHRSI, M. K. e. a. Clustering smart card data for urban mobility analysis. **IEEE Transactions on Intelligent Transportation Systems**, v. 18, p. 712–728, 2017.

MAKRIDAKIS, S. G.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. **Forecasting: Methods and Applications**. 3rd. ed. [S.l.]: Wiley, New York, 1997.

NOVAES, A. G. **Sistemas de Transportes. Volume 1: Análise de Demanda.** [S.l.]: Editora Edgard Blucher Ltda., 1986.

ORTÚZAR, J. D. D.; WILLUMSEN, L. G. **Modelling transport**. 4. ed. [S.l.]: Wiley, New Jersey, 2011.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>.

RIO DE JANEIRO. **Plano diretor de transporte urbano da Região Metropolitana do Rio de Janeiro 2013**. Rio de Janeiro, 2013. Disponível em: <http://thecityfixbrasil.com/files/2015/08/Rio-de-Janeiro_2013.pdf>.

SHEFFI, Y. **Urban Transportation Networks: equilibrium analysis with mathematical programming methods**. [S.l.]: Prentice-Hall, Englewood Cliff, 1985.

SHEPHERD, G. M. **The synaptic organization of the brain**. 5. ed. [S.l.]: Oxford University Press, 2004.

SãO PAULO. **Pesquisa Origem Destino São Paulo 2007**. São Paulo, 2018. Disponível em: <https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino>.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. 1. ed. [S.l.]: Addison Wesley, 2005.

TSUNG-HSIEN, T.; CHI-KANG, L. et al. Neural network based temporal feature models for short-term railway passenger demand forecasting. **Elsevier**, v. 36, p. 3728–3736, 2009.

VUCHIC, V. R. **Urban Transit: operations, planning and economics**. [S.l.]: Wiley New Jersey, 2005.

WARDROP, J. G. Road paper. some theoretical aspects of road traffic research. **Proceedings of the Institution of Civil Engineers**, v. 1, p. 325–362, 1952.

ZHANG, K.; FENG, Z. et al. A framework for passengers demand prediction and recommendation. **IEEE International Conference on Services Computing**, 2016.